Universidad
Carlos III de Madrid

TESIS DOCTORAL

Fiedler matrices: Numerical and structural properties

Autor:
Javier Pérez Álvaro

Directores:
Fernando De Terán y Froilán M. Dopico

Departamento de matemáticas

Leganés, Abril, 2015

Universidad
Carlos III de Madrid

TESIS DOCTORAL

FIEDLER MATRICES: NUMERICAL AND STRUCTURAL PROPERTIES

Autor:

Javier Pérez Álvaro

Directores:

Fernando De Terán y Froilán M. Dopico

Firma del Tribunal Calificador:

		Firma
Presidente:	_____	_____
Vocal:	_____	_____
Secretario:	_____	_____

Calificación:

Leganés, de de 201

Contents

Chapter 1

Motivation, introduction, and summary of main results

The First and second Frobenius companion matrices appear frequently in numerical application, but it is well known that they posses many properties that are undesirable numerically, which limit their use in applications. *Fiedler companion matrices*, or Fiedler matrices for brevity, introduced in 2003, is a family of matrices which includes the two Frobenius matrices. The main goal of this work is to study whether or not Fiedler companion matrices can be used with more reliability than the Frobenius ones in the numerical applications where Frobenius matrices are used. For this reason, in this work we present a thorough study of Fiedler matrices: their *structure and numerical properties*, where we mean by numerical properties those properties that are interesting for applying these matrices in numerical computations, and some of their *applications in the field on numerical linear algebra*.

The introduction of Fiedler companion matrices is an example of a simple idea that has been very influential in the development of several lines of research in the numerical linear algebra field. This family of matrices has important connections with a number of topics of current interest, including: polynomial root finding algorithms, linearizations of matrix polynomials, unitary Hessenberg matrices, CMV matrices, Green's matrices, orthogonal polynomials, rank structured matrices, quasiseparable and semiseparable matrices, etc (for a more detailed survey of the influence of Fiedler companion matrices in numerical linear algebra and matrix analysis see [104]).

In this introductory chapter we will present a brief introduction to Fiedler companion matrices. Then, in order to motivate better the importance of Fiedler matrices and the study of their numerical and structural properties, we will also summarize the areas in which Fiedler matrices may play or are playing a relevant role. Finally, we will present a summary of the chapters presented in this thesis and the main original contributions contained in them.

1.1 A brief introduction to Fiedler companion matrices

Fiedler matrices first appeared in the context of *companion matrices of monic polynomials* in [59]. Since in this work we are going to deal with companion matrices, we present in Definition 1.1 what we mean by a companion matrix of a monic polynomial

$$p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k \quad \text{with } a_k \in \mathbb{C} \text{ for } k = 0, 1, \ldots, n-1. \tag{1.1}$$

This definition is a particular case of the more general definition presented in [48] of a companion form of grade ℓ of a matrix polynomial.

Definition 1.1. *Given a monic polynomial $p(z)$ of degree $n \geq 2$, a* companion matrix *of $p(z)$ is a matrix $A \in \mathbb{C}^{n\times n}$ satisfying the following two properties:*

(i) *Each entry of the matrix A is either a constant $\alpha \in \mathbb{C}$, or a constant times one of the coefficients of $p(z)$, i.e., βa_j for some $\beta \in \mathbb{C}$ and $0 \leq j \leq n-1$, and*

(ii) *the eigenvalues of A are equal to the roots of $p(z)$, or equivalently, the characteristic polynomial of A satisfies $\det(zI - A) = p(z)$.*

The best well known examples of companion matrices of the monic polynomial (1.1) are the matrices

$$C_1 := \begin{bmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_1 & -a_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad C_2 := \begin{bmatrix} -a_{n-1} & 1 & 0 & \cdots & 0 \\ -a_{n-2} & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -a_1 & 0 & 0 & \cdots & 1 \\ -a_0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{1.2}$$

known as the *first* and *second Frobenius companion matrices* of $p(z)$, respectively. Notice that C_1 and C_2 can be constructed directly from the coefficients of the polynomial $p(z)$. The use of companion matrices goes back at least to Frobenius (1879) in his "rational canonical form" of a matrix [64]. Other similar Frobenius companion matrices that appear in the literature can be obtained by transposition and/or by reversing the order of rows and columns of C_1 or C_2 [87, pp 194–200] and [99, p. 105].

Frobenius companion matrices are important in theory, in numerical computations, and in applications. For instance, MATLAB command `roots` computes all the roots of a polynomial by applying the Francis' implicitly-shifted QR-algorithm, or QR-algorithm for short, to a balanced Frobenius companion matrix [113]. Frobenius companion matrices are also widely used in control theory and signal processing, for example, in the observable canonical form as well as the controllable canonical form (see [95] and [99, Section 10.4] and the references therein). However, as we commented before, it is well known that they have many properties that are undesirable numerically.

In 2003, Fiedler expanded significantly the family of companion matrices associated with a monic polynomial [59]. These matrices were named *Fiedler matrices* in [45]. The family of Fiedler matrices includes C_1 and C_2 but, provided that $n \geq 3$, it contains some other different matrices and, in fact, many others when n is large. Every Fiedler matrix shares with C_1 and C_2 two key properties: (i) its characteristic polynomial is $p(z)$ in (1.1), and (ii) $(n-1)$ of its nonzero entries are equal to 1 and the remaining nonzero entries are equal to a_i, for $i = 0, \ldots, n-1$, with exactly one copy of each. In fact, they are easily constructible from the polynomial, without performing any arithmetic operation, by means of a uniform template valid for all polynomials [47, Algorithm 1]. According to Definition 1.1, this justifies that Fiedler matrices are also called Fiedler companion matrices.

The first key observation made in [59] towards the introduction of the family of Fiedler companion matrices is that C_1 and C_2 have a simple factorization. For the monic polynomial $p(z)$ (1.1), if we define the $n \times n$ matrices [1]

$$M_0 := \begin{bmatrix} I_{n-1} & 0 \\ 0 & -a_0 \end{bmatrix} \quad \text{and} \quad M_k := \begin{bmatrix} I_{n-k-1} & & & \\ & -a_k & 1 & \\ & 1 & 0 & \\ & & & I_{k-1} \end{bmatrix}, \quad k = 1, \ldots, n-1, \tag{1.3}$$

[1] Here and in the rest of this work I_j denotes the $j \times j$ identity matrix.

then $C_1 = M_{n-1} \cdots M_1 M_0$ and $C_2 = M_0 M_1 \cdots M_{n-1}$.

The second key idea in [59] was to notice that every matrix that can be obtained multiplying the n matrices (1.3) in any order is similar to the Frobenius companion matrices (1.2). Hence, the matrices (1.3) are the basic factors used to build all Fiedler matrices. In [59], Fiedler matrices are defined as the product

$$M_\sigma = M_{i_1} M_{i_2} \cdots M_{i_n},$$

where $\sigma = (i_1, i_2, \ldots, i_n)$ is any possible permutation of the n-tuple $(0, 1, \ldots, n-1)$. With this notation we can state formally the key result proved by Fiedler.

Theorem 1.2. *Given a monic polynomial $p(z)$, all Fiedler matrices M_σ associated with $p(z)$ are similar to each other.*

Since Frobenius companion matrices are particular examples of Fiedler matrices and the characteristic polynomial of a matrix is invariant under similarity, then, we have that the characteristic polynomial of all associated Fiedler matrices of $p(z)$ is equal to $p(z)$.

The third key observation is that some permutations may produce matrices with interesting structures. Fiedler showed that we could generate a different companion matrix by arranging the coefficients, alternating with zeros, along the super- and subdiagonal, together with ones and zeros along the supersuper- and subsubdiagonal of a pentadiagonal matrix. In order to get this companion matrix, consider the permutation $\tau = (0, 2, 4, \ldots, 1, 3, 5, \ldots)$ with all the even indices gathered together and all the odd indices gathered together. Here is a 10×10 example:

$$\begin{bmatrix} -a_9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -a_8 & 0 & -a_7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_6 & 0 & -a_5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -a_4 & 0 & -a_3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -a_2 & 0 & -a_1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -a_0 & 0 \end{bmatrix}. \tag{1.4}$$

This Fiedler matrix is a *pentadiagonal* matrix for any degree n of the polynomial $p(z)$. For high degree polynomials, this companion matrix will have much lower bandwidth than the Frobenius companion matrices, and may have potential advantages in numerical computations.

In summary, the family of Fiedler companion matrices has expanded significantly the available set of companion matrices beyond the Frobenius ones. Moreover, Frobenius companion matrices appear frequently in the literature on *control and signal processing, polynomial root finding algorithms, and bounds for roots of polynomial*, and, therefore, it is natural to investigate whether other Fiedler matrices are better suited for these applications than the Frobenius companion matrices or not. Also, since the introduction of this family of matrices, the ideas in [59] have strongly influenced the development of several lines of research. These lines of research include *linearization of matrix polynomials*, and *rank structured matrices*. So Fiedler matrices provide opportunities for further developments from both theoretical and numerical points of view.

1.2 Polynomial root-finding using companion matrices

The computation of roots of scalar polynomials, that we term as *the polynomial root finding problem*, is one of the oldest mathematical problems [26, 129]. The solutions of quadratic equations were known to the Arab scholars in the early Middle Ages. The cubic and quartic equations were

solved in closed form in the mid-16th century thanks to the work of the Italian mathematicians S. del Ferro, N. Tartaglia, G. Cardano, and L. Ferrari. However, in the early 19th century, N. H. Abel showed that polynomials of degree five or more can not be solved with a formula involving sums, differences, products, quotients, and radicals of the coefficients. Since then, many researchers have concentrated on numerical iterative methods to solve polynomial equations.

Computing roots of polynomials has important applications in many areas. Low-degree polynomial equations appear in physics, chemistry, engineering, business management and statistics and they are solved using tools of linear algebra [129]. These tools are sufficiently effective in most cases. Polynomial equations with large degree (typically 100 and sometimes of order of several thousands) arise in computer algebra, computational algebraic geometry, control theory and signal processing (see [114, 115, 116, 125, 129] and the bibliography therein), and, until recently, they were a challenge for the available software. Another major application of the polynomial root-finding problem may be found in the symbolic and numerical solution of systems of polynomial equations. Different methods have been proposed that reduce the system of polynomial equations to a single polynomial such that the solution of the former can be obtained from the roots of the latter [139]. The polynomials obtained usually have large degrees and huge coefficients. Finally, we refer the reader to the URL `http://www.elsevier.com/locate/cam` where more than 8000 references about the polynomial root-finding problem are cited.

Nowadays we are aware of many algorithms for computing roots of polynomials, together with their numerical analysis and applications. These algorithms can be classified in two categories. In the first category we have algorithms that are implemented using the standard machine floating point arithmetic. Some of these algorithms are the Madsen–Reid [107], Jenkins–Traub [90], Aberth-Ehrlich [1, 56], Durand-Kerner [96], Laguerre [142], QR-algorithm applied to companion matrices, QZ-algorithm applied to companion pencils (we refer to [117, 118] for thorough surveys on numerical methods for computing the roots of polynomials). In the second category we have algorithms that work using different levels of working precision with an increasing number of bits. In this second category we have the algorithm MPSolve [17, 18], which can compute roots of polynomials with any number of digits of precision.

Among the algorithms that work with a fixed precision, the most widely used are the ones that compute the roots of polynomials as the eigenvalues of companion matrices. The main advantages of this approach are that these methods are easy to implement (given any algorithm to compute the eigenvalues of a matrix), standard backward stable eigenvalue algorithms like the QR-algorithm may be used to compute the eigenvalues of the companion matrix, known a priory backward stability in the matrix sense, and guaranteed convergence in practice. As we commented, this is the approach followed by the MATLAB command `roots` which uses the QR-algorithm on the Frobenius companion matrix (1.2) to get its eigenvalues. Though this may not be the best way to address the polynomial root-finding problem, from the point of view of efficiency and storage (a polynomial has $O(n)$ coefficients, while the complexity of the QR-algorithm is $O(n^2)$ storage and $O(n^3)$ floating point operations, or flops [123]), it has been extensively used because of its robustness and backward stability. Nonetheless, to overcome the mentioned drawbacks on the efficiency (measured in number of operations) and storage, several fast variants of the QR method have been proposed, which take advantage of the structure of the Frobenius companion matrix (see, for instance, [9, 19, 20, 22, 34, 36, 69, 156]), or variants of the LR algorithm [172], but none of them has been proved to be backward stable. In a different line of research, also variants of C_1, C_2 have been proposed, devoted to improve the accuracy in the case of multiple roots, where the standard companion matrix gives less accurate results than for simple roots (see [28, 127]).

Even though computing the roots of a polynomial and computing the eigenvalues of a companion matrix are mathematically equivalent, these two problems present relevant differences from the numerical point of view. In particular, those regarding conditioning and backward errors. The difference in this setting relies on the fact that, due to perturbations, the companion matrix may become a dense matrix, which has not the structure of a companion matrix any more. In

other words, small perturbations of the companion matrix might not correspond to equally small perturbations of the associated polynomial. In [53, 150] the numerical properties of the Frobenius companion matrices (eigenvalue condition numbers and backward errors of computed roots of polynomials using Frobenius companion matrices) were studied. In Section 1.2.1, we summarize the work done in [53] where the backward errors of computed roots of polynomials using Frobenius companion matrices are analyzed, and we emphasize the drawbacks, from the point of view of backward errors, of using Frobenius companion matrices. Then, in Section 1.2.2 we summarize the work carried out in [150] where the authors studied the eigenvalue condition numbers and pseudospectra of Frobenius companion matrices, and we stress, from the point of view of condition numbers and pseudospectra, the drawback of using Frobenius companion matrices. In Section 1.2.3 we explain the concept of balancing a matrix, and we summarize the results found in [53] and [150] when the Frobenius companion matrices are balanced. Finally, in Section 1.2.4 we explain the role that Fiedler companion matrices may play as a new tool for finding the roots of monic polynomials.

1.2.1 Backward stability of polynomial root-finding using Frobenius companion matrices

Suppose that the roots of a monic polynomial $p(z)$ are computed as the eigenvalues of a companion matrix A of $p(z)$ using a backward stable eigenvalue algorithm like the QR-algorithm [10]. The backward stability of the eigenvalue algorithm ensures that the whole set of computed eigenvalues is the whole set of exact eigenvalues of a matrix $A + E$, where E is a dense matrix such that

$$\|E\| = O(u)\|A\|, \tag{1.5}$$

for some matrix norm $\|\cdot\|$, and where u denotes the machine epsilon. However, this does not guarantee that these (computed) eigenvalues are the roots of a nearby polynomial of $p(z)$ or, in other words, that the method is backward stable from the point of view of the polynomials. In order for the method to be backward stable from the point of view of the polynomials, the computed eigenvalues should be the exact roots of a polynomial $\widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k$ satisfying

$$\frac{\|\widetilde{p} - p\|}{\|p\|} = O(u), \tag{1.6}$$

for some polynomial norm $\|\cdot\|$, if we are interested in normwise backward stability or the strongest requirement,

$$\max_{k=0,1,\ldots,n-1} \frac{|\widetilde{a}_k - a_k|}{|a_k|} = O(u), \tag{1.7}$$

if we are interested in coefficientwise backward stability. If (1.6) (resp. (1.7)) holds, the root-finding method using the companion matrix A will be normwise (resp. coefficientwise) backward stable.

To see if (1.6) or (1.7) hold, the key idea is that the computed roots are the exact eigenvalues of a certain perturbation of A, say $A + E$. In other words, we have $p(z) = \det(zI - A)$ and $\widetilde{p}(z) = \det(zI - (A + E))$, with E satisfying (1.5). Therefore, the difference between $p(z)$ and $\widetilde{p}(z)$ can be measured from the variation of the coefficients of the characteristic polynomial of A under small perturbations of A. Hence, the kth coefficient of the characteristic polynomial of a matrix $X = (x_{ij}) \in \mathbb{C}^{n\times n}$ may be considered as a function of the entries of X, $a_k(X) : \mathbb{C}^{n^2} \to \mathbb{C}$, for $k = 0, 1, \ldots, n-1$. Equivalently:

$$\det(zI - X) = z^n + \sum_{k=0}^{n-1} a_k(X) z^k.$$

The function $a_k(X)$ is a multivariable polynomial function of the entries of the matrix X. Therefore, the first order term in E of its Taylor polynomial centered at A is (see, for instance, [74, Th. 3.8] for functions of several complex variables)

$$a_k(A+E) = a_k(A) + \sum_{i,j=1}^{n} \left.\frac{\partial a_k(X)}{\partial x_{ij}}\right|_{X=A} E_{ij} = a_k(A) + \nabla a_k(A) \cdot \mathrm{vec}(E), \tag{1.8}$$

for $k = 0, 1, \ldots, n-1$, where, for a given $m \times n$ matrix $B = (b_{ij})$, $\mathrm{vec}(B)$ is the *vectorization* of B, namely, the column vector

$$\mathrm{vec}(B) := [b_{11}, \ldots, b_{m1}, b_{12}, \ldots, b_{m2}, \ldots, b_{1n}, \ldots, b_{mn}]^T \tag{1.9}$$

(see [87, Def. 4.2.9], for instance), and

$$\nabla a_k(A) = \left[\left.\frac{\partial a_k(X)}{\partial x_{11}}\right|_{X=A} \cdots \left.\frac{\partial a_k(X)}{\partial x_{n1}}\right|_{X=A} \left.\frac{\partial a_k(X)}{\partial x_{12}}\right|_{X=A} \cdots \left.\frac{\partial a_k(X)}{\partial x_{n2}}\right|_{X=A} \cdots \left.\frac{\partial a_k(X)}{\partial x_{1n}}\right|_{X=A} \cdots \left.\frac{\partial a_k(X)}{\partial x_{nn}}\right|_{X=A} \right]. \tag{1.10}$$

Therefore, to first order in E, we have

$$|a_k(A+E) - a_k(A)| = |\nabla a_k(A) \cdot \mathrm{vec}\ (E)|.$$

Hence, the backward stability of the method of computing roots of polynomials using companion matrices may be analyzed by studying the entries of the gradients $\nabla a_k(A)$, for $k = 0, 1, \ldots, n-1$.

For a precedent on the perturbation analysis of the coefficients of the characteristic polynomial of a matrix, we refer the reader to [89]. In that paper, several bounds are derived for the variation of the characteristic polynomial of an arbitrary matrix A under perturbations, in terms of symmetric functions of the singular values of A, but the bounds there are very pessimistic for general matrices. Also, in the recent reference [100], the authors address this problem, namely, to know whether or not solving the polynomial root-finding problem as an eigenvalue problem is backward stable from the point of view of the polynomials, but they use a suitable companion matrix for the polynomial expressed in barycentric form. In that reference the polynomials are not necessarily monic, but the authors follow a similar approach to ours.

When A is one of the Frobenius companion matrices, the backward stability of the polynomial root-finding method using companion matrices was studied in [53]. If we focus on the first Frobenius companion matrix, in [53] it was shown that, if

$$\widetilde{p}(z) = \det(zI - C_1 - E) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k, \tag{1.11}$$

where C_1 is the first Frobenius companion matrix defined in (1.2), then, to first order in (the entries of) E,

$$\widetilde{a}_k - a_k = \sum_{s=0}^{k} \sum_{j=1}^{n-k-1} a_s E_{j-s+k+1,j} - \sum_{s=k+1}^{n} \sum_{j=n-k}^{n} a_s E_{j-s+k+1,j}. \tag{1.12}$$

If the eigenvalues of C_1 are computed with a backward stable algorithm, it may be proved from (1.12) that, to first order in E, the computed eigenvalues are the exact roots of a polynomial $\widetilde{p}(z)$ as in (1.11) such that

$$\frac{\|\widetilde{p} - p\|}{\|p\|} = O(u)\|p\|, \tag{1.13}$$

since E satisfies $\|E\| = O(u)\|C_1\| = O(u)\|p\|$. Note that (1.13) does not imply that computing the roots of $p(z)$ using C_1 (or C_2) is a backward stable method from the point of view of the polynomials, since large values of $\|p\|$ can give large backward errors.

This penalty in the transition from matrix to polynomial backward error is an intrinsic matrix perturbation phenomenon, independent of the algorithm, and it is determined by the particular properties of the Frobenius companion matrix C_1 and the magnitude of $\|E\|$ in (1.11). In fact, one may consider in (1.11) a more general matrix E such that

$$\|E\| = \alpha(p)O(u)\|C_1\| = \alpha(p)O(u)\|p\|, \tag{1.14}$$

with $\alpha(p)$ being some positive quantity depending of $p(z)$. Matrices as in (1.14) are obtained as the matrix backward errors of the fast variants of the QR-algorithm. For example, in the analysis of the algorithm presented in [8] a matrix backward error $\|E\| = O(u)\|C_1\|^2/\min\{1, |a_0|\}$ is obtained, that is, for that algorithm we have that $\alpha(p) = \|C_1\|/\min\{1, |a_0|\}$. Then, using (1.12) and (1.14), we could replace (1.13) by

$$\frac{\|\widetilde{p} - p\|}{\|p\|} = \alpha(p)O(u)\|p\|. \tag{1.15}$$

Equations (1.13) and (1.15) show that, even using backward stable algorithms to compute the eigenvalues of C_1 (or C_2), the approach of computing the roots of a polynomial $p(z)$ as the eigenvalues of its Frobenius companion matrix is not backward stable from the point of view of the polynomials.

One possible way to circumvent large polynomial backward errors due to the occurrence of large polynomial coefficients is to shift from companion matrices to companion pencils. A companion pencil of a (non necessarily monic) polynomial $p(z) = \sum_{k=0}^{n} a_k z^k$ is the set of matrices of the form $zA - B$, with $A, B \in \mathbb{C}^{n\times n}$ and $z \in \mathbb{C}$, such that $\det(zA - B) = p(z)$. For example, the first Frobenius companion pencil associated with the polynomial $p(z) = \sum_{k=0}^{n} a_k z^k$ is

$$z\begin{bmatrix} a_n & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} - \begin{bmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_1 & -a_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \tag{1.16}$$

Then, to compute the roots of the polynomial $p(z)$ one may use any backward stable algorithm, like the QZ-algorithm, to solve the generalized eigenvalue problem $\det(zA - B) = 0$. If one follows this approach, the backward stability of the QZ-algorithm implies that the whole ensemble of computed eigenvalues are the exact eigenvalues of the pencil $z(A + E) - (B + F)$, with $\|E\| = O(u)\|A\|$ and $\|F\| = O(u)\|B\|$ (see [73]). Again, backward stability in the matrix sense does not necessarily imply polynomial backward stability [101]. However, it is shown in [101, 160] that polynomial backward stability can now be achieved by using the QZ-algorithm and the Frobenius companion pencil (1.16), provided that the polynomial $p(z)$ has been previously scaled so that all coefficients have absolute value less than or equal to 1. For polynomials not necessarily monic, this condition can be always achieved by dividing all coefficients of the original polynomial $p(z)$ by some sufficiently large number. This result is also in accordance with the recent results in [126], where the authors study the polynomial backward stability of algorithms that compute the roots of polynomials via the eigenvalues of comrade pencils, where a comrade matrix is the generalization of a companion matrix when, instead of the monomial basis, a orthogonal polynomial basis is used to represent the polynomial. However, this approach has an important drawback from the point of view of storage and efficiency. The storage and number of floating point operations required by the QZ-algorithm is two and three times, respectively, the storage and number of floating point operations required by the QR-algorithm [73]. Nonetheless, to overcome the drawbacks on the efficiency and storage a fast version of the QZ-algorithm has been presented in [25], although no backward error analysis is provided.

1.2.2 Eigenvalue condition numbers and pseudospectra of Frobenius companion matrices

When computing an eigenvalue of a given matrix or a root of a given polynomial it is important to be able to measure their sensitivity to uncertainty in the entries of the matrix or in the coefficients of the polynomial, respectively, since the maximum possible accuracy of the computation is determined by this. Zeros of polynomials and eigenvalues of non normal matrices are well-known examples of problems whose answers may be highly sensitive to perturbations. The sensitivity of these two problems was made famous by Wilkinson in the early 1960s [132, 168, 169]. Both problems are related since the zeros of a monic polynomial are equal to the eigenvalues of any companion matrix of the polynomial.

Much has been done to derive ways to estimate the influence of perturbations on the roots of polynomials and on the eigenvalues of matrices. One approach is to derive a condition number to estimate the largest magnitude of the changes of the roots and eigenvalues which corresponds to changes in the coefficients of the polynomial or in the entries of the matrix when the only information on these changes is their size (norm or absolute value). The other approach is to consider the perturbation as a continuity problem and to use the geometry of the complex plane, that is, the use of the concepts of pseudospectrum of a matrix and pseudozero set of a polynomial. Our interest in pseudospectra of matrices, pseudozero sets of polynomials, condition numbers of roots of polynomials, and eigenvalue condition numbers comes from using them as tools for comparing the sensitivity of the roots of a polynomial with the sensitivity of the eigenvalues of a companion matrix of the polynomial.

In the following two sections we summarize the results obtained in [150] concerning pseudospectra and eigenvalue condition numbers of Frobenius companion matrices. In order to better express these results, we need to distinguish between norms on the vector space of polynomials of degree less than or equal to n and norms on the vector space of coefficients of monic polynomials (excluding the leading coefficient $a_n = 1$) of degree equal to n. In particular, for a polynomial $p(z) = \sum_{k=0}^{n} a_k z^k$ non necessarily monic, $\|p\|_2$ is the norm on the vector space of polynomials of degree less than or equal to n defined as

$$\|p\|_2 = \sqrt{\sum_{k=0}^{n} |a_k|^2},$$

In addition, for a monic polynomial $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$, we define $\|\!|p|\!\|_2$ as

$$\|\!|p|\!\|_2 = \sqrt{\sum_{k=0}^{n-1} |a_k|^2}.$$

Notice that $\|\!|p|\!\|_2$ is not a norm on the vector space of polynomials of degree less than or equal to n. Also notice that given two monic polynomials $p(z)$ and $q(z)$ of degree n, we have that $\|p - q\|_2 = \|\!|p - q|\!\|_2$.

1.2.2.1 Eigenvalue condition numbers of Frobenius companion matrices and condition numbers of roots of monic polynomials

Suppose that the roots of a monic polynomial $p(z)$ are computed as the eigenvalues of a companion matrix A of $p(z)$. Ideally, one would want the eigenvalues of A to be as well conditioned as the roots of $p(z)$. Recall that the condition number of a simple nonzero root λ of the monic polynomial (1.1) is

$$\kappa(\lambda, p) := \lim_{\epsilon \to 0} \sup \left\{ \frac{|\widetilde{\lambda} - \lambda|}{\epsilon |\lambda|} : \widetilde{\lambda} \text{ is a root of } \widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k \text{ with } \|\!|\widetilde{p} - p|\!\|_2 \leq \epsilon \|\!|p|\!\|_2 \right\}. \quad (1.17)$$

Note that the condition number $\kappa(\lambda, p)$ measures the relative sensitivity of the simple root λ with respect to relative normwise perturbations of $p(z)$. If one is interested in coefficientwise relative perturbations, then is more convenient to use the following condition number:

$$\operatorname{cond}(\lambda, p) := \lim_{\epsilon \to 0} \sup \left\{ \frac{|\widetilde{\lambda} - \lambda|}{\epsilon |\lambda|} : \widetilde{\lambda} \text{ is a root of } \widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k \text{ with } \sqrt{\sum_{k=0}^{n-1} \left| \frac{\widetilde{a}_k - a_k}{a_k} \right|^2} \leq \epsilon \right\}. \tag{1.18}$$

The condition number $\operatorname{cond}(\lambda, p)$ measures the relative sensitivity of the simple root λ with respect to relative coefficientwise perturbations of $p(z)$.

The condition numbers $\kappa(\lambda, p)$ and $\operatorname{cond}(\lambda, p)$ in (1.17) and (1.18), respectively, can be computed explicitly (see Section 8.1). With the notation

$$\Lambda(z) = \begin{bmatrix} z^{n-1} & \cdots & z & 1 \end{bmatrix}^T \quad \text{and} \quad \widehat{\Lambda}(z) = \begin{bmatrix} z^{n-1} a_{n-1} & \cdots & z a_1 & a_0 \end{bmatrix}^T, \tag{1.19}$$

we have

$$\kappa(\lambda, p) = \frac{|||p|||_2 \|\Lambda(\lambda)\|_2}{|\lambda| \cdot |p'(\lambda)|} \quad \text{and} \quad \operatorname{cond}(\lambda, p) = \frac{\|\widehat{\Lambda}(\lambda)\|_2}{|\lambda| \cdot |p'(\lambda)|}. \tag{1.20}$$

From (1.20) is easy to prove that $\operatorname{cond}(\lambda, p) \leq \kappa(\lambda, p)$, and in some situation $\operatorname{cond}(\lambda, p)$ can be much smaller that $\kappa(\lambda, p)$.

The condition number of a simple nonzero eigenvalue λ of a matrix $A \in \mathbb{C}^{n \times n}$ is

$$\kappa(\lambda, A) := \lim_{\epsilon \to 0} \sup \left\{ \frac{|\widetilde{\lambda} - \lambda|}{\epsilon |\lambda|} : \widetilde{\lambda} \text{ is an eigenvalue of } A + E \text{ with } \|E\|_2 \leq \epsilon \|A\|_2 \right\}. \tag{1.21}$$

This condition number, which measures the relative sensitivity of the simple eigenvalue λ with respect to relative normwise perturbations of A, was introduced in [16] and it is a slight modification of the Wilkinson condition number [169], which measures the absolute sensitivity of a simple eigenvalue with respect to absolute normwise perturbations of the matrix.

The condition number $\kappa(\lambda, A)$ can be computed explicitly (see [73])

$$\kappa(\lambda, A) = \frac{\|x\|_2 \|y\|_2}{|y^T x|} \frac{\|A\|_2}{|\lambda|},$$

where $x, y \in \mathbb{C}^n$ are the right and left eigenvectors of A, respectively, associated with the simple eigenvalue λ.

Remark 1.3. In this work, the right and left eigenvectors of a matrix $A \in \mathbb{C}^{n \times n}$ associated with the eigenvalue λ are two nonzero vectors $x, y \in \mathbb{C}^n$, respectively, that satisfy $Ax = \lambda x$ and $y^T A = \lambda y^T$. The definition of left eigenvector looks nonstandard, since the usual definition of a left eigenvector of a matrix A associated with the eigenvalue λ is a nonzero vector $y \in \mathbb{C}^n$ such that $y^* A = \lambda y^*$ but it will be more convenient in Chapter 8.

When the matrix A is one of the Frobenius companion matrices associated with a monic polynomial $p(z)$, the condition number $\kappa(\lambda, C_i)$, for $i = 1, 2$, can be written in closed-form. With the notation

$$\Pi(z) = [p_0(z), p_1(z), \ldots, p_{n-1}(z)]^T, \tag{1.22}$$

where, for $k = 0, 1, \ldots, n-1$, $p_k(\lambda)$ is the degree k Horner shift of $p(z)$ (see Definition 2.13), the expression for the condition number $\kappa(\lambda, C_i)$, for $i = 1, 2$, then reduces to the following formula (see [150, Proposition 3.2])

$$\kappa(\lambda, C_i) = \frac{\|C_i\|_2}{|\lambda|} \frac{\|\Lambda(\lambda)\|_2 \|\Pi(\lambda)\|_2}{|p'(\lambda)|}, \quad \text{for } i = 1, 2,$$

where we have used that $p'(\lambda) = \sum_{k=0}^{n-1} \lambda^k p_k(\lambda)$.

If one wants to compute the roots of the polynomial $p(z)$ via the eigenvalues of a companion matrix A of the polynomial $p(z)$, one would need the condition number of λ as an eigenvalue of A and the condition number of λ as a root of $p(z)$ to be of the same order of magnitude, that is,

$$\frac{\kappa(\lambda, A)}{\kappa(\lambda, p)} = O(1), \tag{1.23}$$

or the strongest condition,

$$\frac{\kappa(\lambda, A)}{\text{cond}(\lambda, p)} = O(1), \tag{1.24}$$

When A is one of the Frobenius companion matrices, the ratios (1.23) and (1.24) satisfy

$$\frac{\kappa(\lambda, C_i)}{\text{cond}(\lambda, p)} \geq \frac{\kappa(\lambda, C_i)}{\kappa(\lambda, p)} = \frac{\|C_i\|_2}{|||p|||_2} \|\Pi(\lambda)\|_2, \quad \text{for } i = 1, 2.$$

These ratios were studied in [150], and it was shown that when $|||p|||_2 \gg 1$ they can be very large and, so, in general, (1.23) and (1.24) do not hold. Therefore, although in principle one can find the zeros of a polynomial via Frobenius companion matrices, it may not be a good idea to do so if the polynomial has large coefficients.

1.2.2.2 Pseudozero sets of polynomials and pseudospectra of companion matrices

Pseudospectra is a tool that has become popular since the 1990s in the study of matrices and linear operators. This concept extends the traditional notion of eigenvalues or spectra (for matrices or linear operators, respectively) and is an established tool for gaining insight into the sensitivity of the eigenvalues of a matrix to perturbations. Pseudospectra can reveal information about the behavior of systems, both linear and nonlinear, including stability [78, 135, 136], and convergence of matrix iterations, and their use is widespread with applications in areas such as fluid mechanics, hydrodynamics stability and turbulence [134, 140, 152, 153], Markov chains [92], and control theory [84, 85, 86]. For a survey on pseudospectra of square matrices we refer to Trefethen and Embree's book [155]. Also, the notion of pseudospectrum has been extended for matrix pencils [63, 137, 162] and matrix polynomials [77, 148].

For $\epsilon > 0$, the *ϵ-pseudospectrum* of a matrix $A \in \mathbb{C}^{n\times n}$, denoted by $\Lambda_\epsilon(A)$, is the following set in the complex plane (see [150]):

$$\Lambda_\epsilon(A) := \{z \in \mathbb{C} : \ z \text{ is an eigenvalue of } A + E \text{ for some } E \text{ with } \|E\|_2 \leq \epsilon \|A\|_2\}.$$

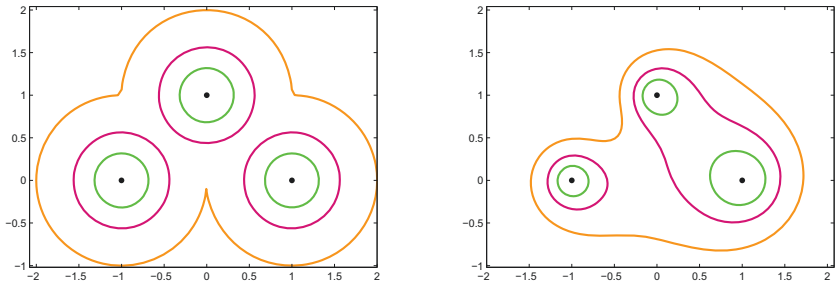
In words, the ϵ-pseudospectrum is the set of numbers that are eigenvalues of some perturbed matrix $A + E$ with $\|E\|_2 \leq \epsilon\|A\|_2$.

Pseudospectra of normal matrices are "uninteresting" sets since the ϵ-pseudospectrum of a normal matrix is the union of open balls of radii ϵ centered at its eigenvalues [155, Theorem 2.2]. Pseudospectra becomes interesting for matrices that are far from being normal. We illustrate this in Figure 1.2.1, where we plot the ϵ-pseudospectra of two 3×3 matrices, being the first one normal and the second one nonnormal, for three different values of ϵ.

Remark 1.4. All the computations of pseudospectra are performed using `MATLAB` and the toolbox `EigTool`. This toolbox is a free package for computing pseudospectra of dense and sparse matrices [170].

Pseudospectra are nontrivial to compute. Algorithms for computing pseudospectra are based on the following characterization of $\Lambda_\epsilon(A)$ in terms of the resolvent $(zI - A)^{-1}$ or in terms of the minimum singular value of $(zI - A)$ [155, Theorem 2.2]:

$$\Lambda_\epsilon(A) = \{z \in \mathbb{C} : \ \|(zI - A)^{-1}\|_2 \geq (\epsilon\|A\|_2)^{-1}\} = \{z \in \mathbb{C} : \ \sigma_{\min}(zI - A) \leq \epsilon\|A\|_2\}, \tag{1.25}$$



(a) normal (b) nonnormal

Figure 1.2.1: The geometry of pseudospectra: schematic view. In each plot the contours represent the boundary of $\Lambda_\epsilon(A)$ for three values of ϵ.

where by convention $\|(zI-A)^{-1}\|_2$ takes the value ∞ in the spectrum of A (where the resolvent is not definded), and where $\sigma_{\min}(zI-A)$ denotes the minimum singular value of $zI-A$. In words, the ϵ-pseudospectrum is the subset of the complex plane bounded by the $(\epsilon\|A\|_2)^{-1}$ level curve of the norm of the resolvent or the $\epsilon\|A\|_2$ level curve of the function $\sigma_{\min}(zI-A)$.

Most algorithms for computing pseudospectra use (1.25) and fall into two broad classes: grid algorithms [68, 98] and path-following algorithms [12, 13, 121]. Path-following algorithms consist of finding a point on the boundary of the desired pseudospectrum and follow from it a curve in the complex plane on which the minimum singular value of $zI-A$ is constant. Grid algorithms compute $\sigma_{\min}(zI-A)$ via the SVD on a regular grid of points in the complex plane and, then, visualizes the data, typically via a contour plot. The obvious grid algorithm is evaluate $\sigma_{\min}(zI-A)$ on a grid in the complex plane and then generate a contour plot from this data. The problem with this algorithm is that computing the SVD of a $n\times n$ matrix on a $m\times m$ grid requires $O(m^2n^3)$ floating point operations which is highly expensive. Considerable progress has been made in making this process as efficient as possible. These methods for speeding up the computations are based on avoiding uninteresting regions of the complex plane [68], and in first reducing A to Hessenberg or triangular form before computing the minimum singular value of $zI-A$ and, then, applying an iterative method to compute the minimum singular value such as the inverse iteration or the inverse Lanczos iteration [103]. With these techniques the overall complexity can be reduced to $O(n^3+n^2m^2)$, an improvement that makes an enormous difference in practice. For a detailed tutorial survey of computational techniques for computing pseudospectra see [154].

For polynomials, the concept equivalent to the $\epsilon-$pseudospectrum of a matrix is the concept of the $\epsilon-$*pseudozero set*, introduced in [124]. For $\epsilon>0$, the $\epsilon-$*pseudozero set* of the polynomial (1.1), denoted by $Z_\epsilon(p)$, is the following set in the complex plane

$$Z_\epsilon(p)=\left\{z\in\mathbb{C}: z \text{ is a root of } \widetilde{p}(z)=z^n+\sum_{k=0}^{n-1}\widetilde{a}_k z^k \quad \text{with} \quad |||\widetilde{p}-p|||_2\le\epsilon|||p|||_2\right\}.$$

If one is interested in coefficientwise relative perturbations, then it is more convenient to use as the $\epsilon-$pseudozero set the following set in the complex plane

$$\text{Pseudo}_\epsilon(p)=\left\{z\in\mathbb{C}: z \text{ is a root of } \widetilde{p}(z)=z^n+\sum_{k=0}^{n-1}\widetilde{a}_k z^k \quad \text{with} \quad \sqrt{\sum_{k=0}^{n-1}|\widetilde{a}_k-a_k|^2\cdot|d_k|^2}\le\epsilon\right\}, \tag{1.26}$$

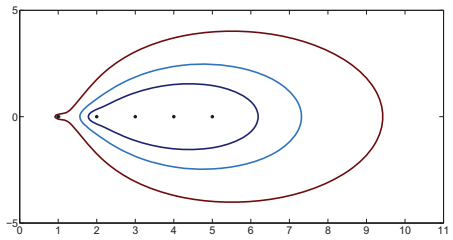
where $d_k=1/|a_k|$ if $a_k\neq 0$ or $d_k=0$ if $a_k=0$, for $k=0,1,\ldots,n-1$.

The pseudozero sets $Z_\epsilon(p)$ and $\text{Pseudo}_\epsilon(p)$ can be characterized in terms of the level curves of certain functions [150, Proposition 2.1]:

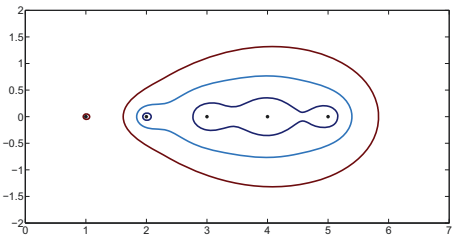
$$Z_\epsilon(p) = \left\{ z \in \mathbb{C} : \quad \frac{|p(z)|}{|||p|||_2 \|\Lambda(z)\|_2} \leq \epsilon \right\} \quad \text{and} \quad \text{Pseudo}_\epsilon(p) = \left\{ z \in \mathbb{C} : \quad \frac{|p(z)|}{\|\widehat{\Lambda}(z)\|_2} \leq \epsilon \right\},$$

where $\Lambda(z)$ and $\widehat{\Lambda}(z)$ are defined in (1.19). This allows to determine $Z_\epsilon(p)$ and $\text{Pseudo}_\epsilon(p)$ numerically using grid algorithms, since computing $|p(z)|/|||p|||_2\|\Lambda(z)\|_2$ and $|p(z)|/\|\widehat{\Lambda}(z)\|_2$ on a $m \times m$ grid requires just $O(nm^2)$ floating point operations. Also, these characterizations of pseudozero sets may be used to prove that $\text{Pseudo}_\epsilon(p) \subseteq Z_\epsilon(p)$.

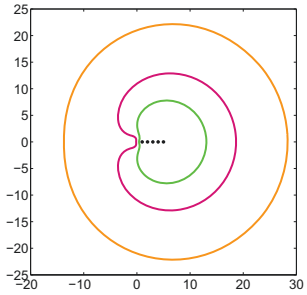
Pseudozero sets of a polynomial can be used to visualize the sensitivity of its roots to perturbations of its coefficients. As we already mentioned before, pseudospectra of Frobenius companion matrices and pseudozero sets of monic polynomials were studied in [150]. The authors found that the pseudozero sets of a monic polynomial $p(z)$ and the pseudospectra of the associated Frobenius companion matrices may be very different, showing that small perturbations of the companion matrix might not correspond to equally small perturbation of the polynomial. We illustrate this in Figure 1.2.2. For $\epsilon = 10^{-3.5}, 10^{-3}, 10^{-2.5}$, we plot in Figure 1.2.2-(a),-(b), and -(c), respectively, $Z_\epsilon(p)$, $\text{Pseudo}_\epsilon(p)$, and the $\epsilon-$pseudospectrum $\Lambda_\epsilon(C_2)$ of the second Frobenius companion matrix of $p(z)$, where $p(z)$ is the monic polynomial whose roots are equal to 1,2,3,4,5. Notice that even for a polynomial of moderate degree ($n = 5$) a big difference between $Z_\epsilon(p)$ and $\text{Pseudo}_\epsilon(p)$, and $\Lambda_\epsilon(C_2)$ can be appreciated.



(a) $Z_\epsilon(p)$



(b) $\text{Pseudo}_\epsilon(p)$



(c) $\Lambda_\epsilon(C_2)$

Figure 1.2.2: For $\epsilon = 10^{-3.5}, 10^{-3}, 10^{-2.5}$, we plot $Z_\epsilon(p)$, $\text{Pseudo}_\epsilon(p)$, and $\Lambda_\epsilon(C_2)$, where $p(z)$ is the monic polynomial whose roots are equal to 1,2,3,4,5.

1.2.3 Balancing Frobenius companion matrices

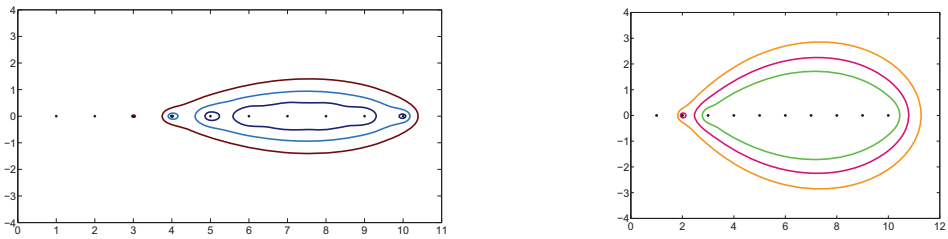
Numerical algorithms that compute the eigenvalues of a nonsymmetric matrix A are typically affected by backward roundoff errors of size roughly $O(u)\|A\|_F$ [128], where u is the machine epsilon. Balancing, an idea introduced in [130], is a standard technique for computing the eigenvalues of a given matrix A, which leads, very often, to more accurate results, especially when the entries of A have very different magnitudes (although there are situations where balancing has the opposite effect [166]). Actually, balancing is implemented by default as an initial step in the MATLAB command `eig` for computing the eigenvalues of arbitrary matrices. Balancing consists of performing a diagonal similarity DAD^{-1} (i.e., with D diagonal) in order to reduce the norm of A by equilibrating as much as possible the ∞-norm of all rows and columns. In addition, very frequently balancing reduces the eigenvalue condition numbers [73, §7.2.2]. Balancing is done with matrices D whose entries are powers of two, so as not to introduce any roundoff errors.

In [53, 150], the authors studied the effect of balancing Frobenius companion matrices on eigenvalue condition numbers, pseudospectra, and backward errors of computed roots of monic polynomials via eigenvalues of Frobenius companion matrices. In this section we denote by $\widetilde{C}_1$ the first Frobenius companion matrix C_1 after being balanced.

In [150] the authors provided some numerical evidence to show that

$$\frac{\kappa(\lambda, \widetilde{C}_1)}{\mathrm{cond}(\lambda, p)} = O(1) \tag{1.27}$$

usually holds, and to show that $\mathrm{Pseudo}_\epsilon(p)$ and $\Lambda_\epsilon(\widetilde{C}_1)$ *usually* are quite close to each other. That is, relative normwise perturbations of the balanced Frobenius companion matrix *tend to be* equivalent to relative coefficientwise perturbations of $p(z)$. This is illustrated in Figure 1.2.3, where, for the polynomial $p(z) = \prod_{i=1}^{10}(z-i)$ and for $\epsilon = 10^{-7}, 10^{-6.5}$, and 10^{-6}, we plot the ϵ-pseudozero sets corresponding to coefficientwise perturbations of the polynomial $p(z)$, and the ϵ-pseudospectra of the balanced Frobenius companion matrix of that polynomial. This agreement of pseudozero



(a) $\mathrm{Pseudo}_\epsilon(p)$ (b) $\Lambda_\epsilon(\widetilde{C}_1)$

Figure 1.2.3: For $\epsilon = 10^{-7}, 10^{-6.5}, 10^{-6}$, we plot $\mathrm{Pseudo}_\epsilon(p)$ and $\Lambda_\epsilon(\widetilde{C}_1)$, where $p(z)$ is the monic polynomial whose roots are equal to $1, 2, \ldots, 10$, and where $\widetilde{C}_1$ denotes the first Frobenius companion matrix of $p(z)$ after being balanced.

sets and pseudospectra suggests that it may be possible to compute roots of monic polynomials stably via eigenvalues of balanced Frobenius companion matrices. Also, in [53] the authors provided some numerical evidence of the fact that finding roots of polynomials using balanced Frobenius companion matrices is a coefficientwise backward stable polynomial root-finding method, that is, (1.7) usually holds.

These facts provide numerical evidence suggesting that *in most cases* the polynomial root-finding problem and the balanced companion matrix eigenproblem are comparable in terms of sensitivity to perturbations and backward stability. Therefore, in these cases, computing roots of

a monic polynomial via eigenvalues of the associated Frobenius companion matrix is numerically reliable, provided that the Frobenius matrix has been balanced. Hence, the command `roots` in MATLAB is usually reliable in a practical sense, although with the drawback of the complexity $O(n^3)$ flops and $O(n^2)$ storage.

We conclude this section with a comment about the stability of computing the roots of a monic polynomial via the eigenvalues of its balanced Frobenius companion matrix. Despite the evidence presented in [53] and [150], the studies carried out in this thesis show that the ratios in (1.6) and in (1.27) may grow with $\|p\|$ moderately, and, so, this approach is not always stable.

1.2.4 Polynomial root-finding using Fiedler companion matrices

As we have seen in Sections 1.2.1 and 1.2.2, the approach of computing the roots of a polynomial $p(z)$ via the eigenvalues of Frobenius companion matrices (1.2) is not stable when $\|p\| \gg 1$. Also, despite the evidence presented in [53, 150], balancing the Frobenius companion matrices is not always enough to guarantee backward stability in the sense (1.6) (see Chapter 9). Moreover, as we have commented, the use of Frobenius companion matrices can be criticized from the point of view of efficiency and storage. A polynomial has $O(n)$ coefficients, while the complexity of the matrix eigenvalue problem is $O(n^2)$ storage and $O(n^3)$ floating point operations, which is certainly too much.

Fiedler companion matrices have expanded significantly the available set of companion matrices beyond the Frobenius ones, and, so, these matrices provide a new tool that could be used instead of the Frobenius companion matrices C_1 and C_2 for computing the roots of a polynomial $p(z)$. For this reason, the eigenvalue condition numbers of Fiedler companion matrices and the backward errors of computed roots using Fiedler matrices are particularly relevant, and the considerations in the previous paragraph give a strong motivation for studying them. In addition, Fiedler matrices like the matrix (1.4) could be used to overcome the mentioned drawbacks on the efficiency and storage. If a numerically reliable matrix eigenvalue algorithm that works on nonsymmetric pentadiagonal matrices in $O(n)$ storage and $O(n^2)$ flops could be found, then the Frobenius companion matrix may be replaced by (1.4) in the polynomial root-finding method using companion matrices.

1.3 Condition numbers for inversion of Frobenius companion matrices

Frobenius companion matrices arise in control theory, particularly, in the study of time-invariant linear systems. For simplicity, we consider here the single-input case:

$$\begin{aligned} &\frac{dx(t)}{dt} = Ax(t) + bu(t), \quad x(0) = x_0 \\ &u(t) = -f^T x(t), \end{aligned}$$

where $A \in \mathbb{C}^{n\times n}$, $b, f \in \mathbb{C}^n$, $u(t) \in \mathbb{C}$, and $x(t) \in \mathbb{C}^n$. The vector f is called the *feedback gain vector*, the equation

$$\frac{dx(t)}{dt} = Ax(t) - bf^T x(t), \quad x(0) = x_0,$$

is called the *closed-loop system*, and its solution is given by $x(t) = e^{(A-bf^T)t}x_0$.

One of the most studied problems for such systems in the *pole placement problem*: Given a set of n complex numbers $\{\lambda_1, \ldots, \lambda_n\}$, find a vector $f \in \mathbb{C}^n$ such that the set of eigenvalues of $A - bf^T$ is equal to $\{\lambda_1, \ldots, \lambda_n\}$. It is well known [171] that a feedback gain vector f exists for all sets $\{\lambda_1, \ldots, \lambda_n\}$ if and only if (A, b) in *controllable*. We recall that the pair (A, b) is controllable if the rank of the matrix $[b, Ab, A^2b, \ldots, A^{n-1}b]$ is equal to n.

The goal of the pole placement problem is that the implemented poles of the closed loop system should be close to the desired ones. Explicit formulas for the solution of the pole placement problem were introduced in [2, 27] and several numerical algorithms for computing the feedback gain vector f have been developed [39, 122, 133, 163]. Also the perturbation theory of this problem has been studied [120, 146], as well as the stability of the numerical algorithms [37, 38]. If the desired poles of the exact closed loop system are very sensitive to perturbations the goal of the pole placement problem cannot be guaranteed. The perturbation analysis made in [75] shows that there are three ingredients controlling this sensitivity: the norm of the feedback vector f, the spectral condition number of the *closed loop matrix* $A-bf^T$, and the distance to uncontrollability. In [75], it is shown that in general one cannot expect that the closed loop system has a spectrum close to the desired one, if at least one of the three contributing ingredients is large.

Frobenius companion matrices (1.2) arise in control theory because any single-input controllable system can be transformed into a companion form system and, also, because the structure of companion systems greatly simplifies theoretical considerations such as the feedback analysis [95]. However, as n increases, Frobenius companion matrices are known to posses many properties that are undesirable numerically. For instance, stable ones are nearly unstable, controllable ones are nearly uncontrollable, and nonsingular ones are nearly singular, that is, they have *large condition numbers for inversion* $\kappa(C_i)$, for $i = 1, 2$, where

$$\kappa(C_i) = \|C_i\| \cdot \|C_i^{-1}\|, \quad \text{for } i = 1, 2. \tag{1.28}$$

These undesirable numerical properties can even arise in algorithms designed specifically for systems in companion forms [24, 131].

The properties mentioned in the paragraph above for the Frobenius companion matrices were studied in detail in [95]. In particular, the authors studied the behavior of the spectral condition number for inversion $\kappa_2(C_1) = \|C_1\|_2\|C_1^{-1}\|_2$. This condition number is related to the relative distance of C_1 to singularity, and, also, it is useful in establishing bounds on the nearness to instability of systems in companion forms. The analysis of $\kappa_2(C_1)$ presented in [95] is based on the following remarkable property of C_1: it is possible to derive explicit expressions for its singular values and at least $n-2$ of the singular values of C_1 are equal to 1 (see also [99, Section 10.4]). Hence, if C_1 is the first Frobenius companion matrix of the monic polynomial (1.1) and if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ denote its singular values, then $\sigma_2 = \sigma_3 = \cdots = \sigma_{n-1} = 1$, and the largest and the smallest singular values σ_1 and σ_n are the roots of the following explicit expressions:

$$\frac{1 + \sum_{k=0}^{n-1} |a_k|^2 \pm \sqrt{\left(1 + \sum_{k=0}^{n-1} |a_k|^2\right)^2 - 4|a_0|^2}}{2}. \tag{1.29}$$

From (1.29), it could be proved that (see [95])

$$\kappa_2(C_1) = \frac{\sigma_1}{\sigma_n} = \frac{1 + \sum_{k=0}^{n-1} |a_k|^2 + \sqrt{\left(1 + \sum_{k=0}^{n-1} |a_k|^2\right)^2 - 4|a_0|^2}}{2|a_0|}, \tag{1.30}$$

and, from (1.30), the following crude lower and upper bound on $\kappa_2(C_1)$ can be established

$$\frac{1 + \sum_{k=0}^{n-1} |a_k|^2}{2|a_0|} \leq \kappa_2(C_1) \leq \frac{1 + \sum_{k=0}^{n-1} |a_k|^2}{|a_0|}.$$

In plain words, these bounds show that $\kappa_2(C_1)$ is large if and only if $|a_0|$ is small or $|a_i|$ is large for some $i = 0, 1, \ldots, n-1$ (or both). This observation is the main reason behind the results presented in [95]. These results establish than, when n is large, controllable canonical systems

generally have such bad numerical properties that they are essentially useless. Therefore, since any Fiedler companion matrix may replace the Frobenius companion matrix C_1 in the companion form system, it is natural to investigate whether other Fiedler matrices are better conditioned than the Frobenius companion matrices or not.

1.4 Bounds for roots of polynomials using Frobenius companion matrices

Frobenius companion matrices have been used extensively for getting upper and lower bounds for the roots of a monic polynomial $p(z)$. To locate approximately the roots of $p(z)$ just through simple operations on its coefficients is a classical problem that has produced a considerable amount of literature (see the comprehensive surveys [108, 141] and the references therein). Simple location rules are used for theoretical purposes, as establishing sufficient conditions to guarantee that $p(z)$ is stable or that all its roots are inside the unit circle, and they are also used in iterative algorithms for computing the roots of $p(z)$ to find initial guesses of the roots for starting the iteration [17, 18].

Let us denote by λ any root of $p(z)$. Our goal is to find nonnegative numbers $L(p)$ and $U(p)$ depending on the coefficients of $p(z)$, such that

$$L(p) \leq |\lambda| \leq U(p). \tag{1.31}$$

Frobenius companion matrices have been widely used to obtain classic bounds of type (1.31) [87, pp. 316–319], as well as other types of location results for roots of polynomials [119].

When $a_0 \neq 0$, i.e, when $\lambda = 0$ is not a root of $p(z)$ in (1.1), the *monic reversal polynomial of* $p(z)$ [87, p. 318] plays an important role in getting bounds for the roots of $p(z)$. It is defined as follows:

$$p^\sharp(z) := \frac{z^n}{a_0} p(z^{-1}) = z^n + \frac{a_1}{a_0} z^{n-1} + \frac{a_2}{a_0} z^{n-2} + \cdots + \frac{a_{n-1}}{a_0} z + \frac{1}{a_0}.$$

Observe that the roots of $p^\sharp(z)$ are the reciprocals of the roots of $p(z)$. Therefore, the eigenvalues of the Frobenius companion forms of $p^\sharp(z)$, i.e., $C_1(p^\sharp)$ and $C_2(p^\sharp)$ (in this section, we indicate explicitly the dependence of C_1 and C_2 on a certain polynomial $q(z)$ using the notation $C_1(q)$ and $C_2(q)$), are also the reciprocals of the roots of $p(z)$. This can be combined with a well known property of any *family of submultiplicative matrix norms*, i.e., a family of matrix norms $\|\cdot\|$ satisfying $\|AB\| \leq \|A\|\|B\|$ for all $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times p}$ [87, Chapter 5]. This property establishes that if $X \in \mathbb{C}^{n\times n}$ and μ is any eigenvalue of X, then $|\mu| \leq \|X\|$ [87, p. 297] and it can be applied to both $C_i(p)$ and $C_i(p^\sharp)$, for $i = 1, 2$, to prove that

$$\left(\|C_i(p^\sharp)\|\right)^{-1} \leq |\lambda| \leq \|C_i(p)\|, \qquad i = 1, 2, \tag{1.32}$$

for any root λ of $p(z)$, which allows us to get bounds of type (1.31). In practice, (1.32) is only used with the 1-, 2-, ∞-, and Frobenius norms. For a matrix $A = (a_{ij}) \in \mathbb{C}^{m\times n}$, these norms are defined as [79, p. 108]

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_2 = \sigma_{\max}(A), \quad \|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|, \quad \|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2},$$

where $\sigma_{\max}(A)$ denotes the largest singular value of A. Note that $\|A\|_1 = \|A^T\|_\infty$, $\|A\|_2 = \|A^T\|_2$, and $\|A\|_F = \|A^T\|_F$. In [87, pp. 316-318], the inequalities (1.32) are used with $C_2(p)$ and $C_2(p^\sharp)$ and the ∞-, 1-, 2-, and Frobenius norms to get the following classical bounds.

Theorem 1.5. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with complex coefficients and λ be any root of $p(z)$. Then $|\lambda|$ satisfies the following inequalities.*

1. *Cauchy's lower and upper bounds* (*coming from* C_2 *and* $\|\cdot\|_\infty$)*:*

$$\frac{|a_0|}{\max\{1,|a_0|+|a_1|,|a_0|+|a_2|,\ldots,|a_0|+|a_{n-1}|\}} \leq |\lambda| \leq \max\{|a_0|,1+|a_1|,\ldots,1+|a_{n-1}|\}\,.$$

2. *Montel's lower and upper bounds* (*coming from* C_2 *and* $\|\cdot\|_1$)*:*

$$\frac{|a_0|}{\max\{|a_0|,1+|a_1|+|a_2|+\cdots+|a_{n-1}|\}} \leq |\lambda| \leq \max\{1,|a_0|+|a_1|+\cdots+|a_{n-1}|\}\,.$$

3. *Carmichael-Mason's lower and upper bounds* (*coming from* C_2 *and* $\|\cdot\|_2$)*:*

$$\frac{|a_0|}{\sqrt{1+|a_0|^2+|a_1|^2+\cdots+|a_{n-1}|^2}} \leq |\lambda| \leq \sqrt{1+|a_0|^2+|a_1|^2+\cdots+|a_{n-1}|^2}\,.$$

4. *Frobenius' lower and upper bounds* (*coming from* C_2 *and* $\|\cdot\|_F$)*:*

$$\frac{|a_0|}{\sqrt{1+(n-1)|a_0|^2+|a_1|^2+\cdots+|a_{n-1}|^2}} \leq |\lambda| \leq \sqrt{(n-1)+|a_0|^2+\cdots+|a_{n-1}|^2}.$$

Note that if $C_1(p)$ and $C_1(p^\sharp)$ are used instead of $C_2(p)$ and $C_2(p^\sharp)$, then the same bounds are obtained, but Cauchy's bounds are obtained for $\|\cdot\|_1$ and Montel's bounds for $\|\cdot\|_\infty$. It is clear that Carmichael-Mason's bounds are always sharper than Frobenius' bounds, but which are the sharpest among the other bounds depends on the particular polynomial that is considered. However, Cauchy's bounds are essentially the sharpest ones in Theorem 1.5. To be precise, if $U_C(p), U_M(p)$, and $U_{CM}(p)$ denote, respectively, the upper Cauchy's, Montel's, and Carmichael-Mason's bounds, then it is easy to prove that $U_C(p) \leq 2\,U_M(p)$ and $U_C(p) \leq \sqrt{2}\,U_{CM}(p)$ for any $p(z)$. Moreover, if $L_C(p), L_M(p)$, and $L_{CM}(p)$ denote, respectively, the lower Cauchy's, Montel's, and Carmichael-Mason's bounds, then $L_M(p) \leq 2\,L_C(p)$ and $L_{CM}(p) \leq \sqrt{2}\,L_C(p)$ for any $p(z)$.

The bounds in Theorem 1.5 have an important drawback: the lower bounds are always smaller than 1 and the upper bounds are always larger than 1. This is a consequence of the presence of entries equal to 1 in the Frobenius companion matrix. For $C_1(p)$ and $C_2(p)$ an standard way to overcome this drawback is to use diagonal similarities, which do not change neither the eigenvalues nor the zero pattern, and to use (1.32). More precisely, let D and $\widetilde{D}$ be nonsingular diagonal matrices, then from (1.32) we get $\left(\|\widetilde{D}^{-1}C_i(p^\sharp)\widetilde{D}\|\right)^{-1} \leq |\lambda| \leq \|D^{-1}C_i(p)D\|$, for $i=1,2$. Given a polynomial $p(z)$, the selection of a proper D (and/or $\widetilde{D}$) may improve drastically the bounds, but a choice of D that is good for certain polynomials may be bad for others, so the choice of proper diagonal similarities is not immediate. Some specific D's have been used to get the well-know Fujiwara's bounds [65]

$$\frac{1}{2\max\left\{\left|\frac{a_1}{a_0}\right|,\left|\frac{a_2}{a_0}\right|^{1/2},\ldots,\left|\frac{a_{n-1}}{a_0}\right|^{1/(n-1)},\left|\frac{1}{2a_0}\right|^{1/n}\right\}} \leq |\lambda| \leq 2\max\left\{|a_{n-1}|,|a_{n-2}|^{1/2},\ldots,|a_1|^{1/(n-1)},\left|\frac{a_0}{2}\right|^{1/n}\right\},$$

as well as Kojima's bounds [97] (see also [87, p. 319])

$$\frac{1}{2\max\left\{\left|\frac{a_1}{a_0}\right|,\left|\frac{a_2}{a_1}\right|,\ldots,\left|\frac{a_{n-1}}{a_{n-2}}\right|,\left|\frac{1}{2a_{n-1}}\right|\right\}} \leq |\lambda| \leq 2\max\left\{|a_{n-1}|,\left|\frac{a_{n-2}}{a_{n-1}}\right|,\ldots,\left|\frac{a_1}{a_2}\right|,\left|\frac{a_0}{2a_1}\right|\right\}.$$

Apart from their eigenvalues, all Fiedler matrices of $p(z)$ share a key property with the first and second Frobenius companion forms: they contain, in different positions, exactly the same nonzero entries, i.e., $n-1$ entries equal to 1, and n entries equal to $-a_0, -a_1, \ldots, -a_{n-1}$. Therefore, it is natural to investigate if matrix norms of Fiedler companion matrices may be used to obtain new and sharper lower and upper bounds on the roots of monic polynomials.

1.5 Linearizations of matrix polynomials

Although matrix polynonials are not covered in this work, Fiedler matrices are becoming an interesting tool in this field. An $m \times n$ matrix polynomial $P(\lambda)$ is a polynomial in λ whose coefficients are $m \times n$ matrices:

$$P(\lambda) = \sum_{i=0}^{k} A_i \lambda^i, \quad A_0, \ldots, A_k \in \mathbb{C}^{m \times n}, \quad A_k \neq 0,$$

where k is the *degree* of $P(\lambda)$. Equivalently, a matrix polynomial is an $m \times n$ matrix whose entries are polynomials in λ. It is said that $P(\lambda)$ is *regular* if $n = m$ and the determinant of $P(\lambda)$ is not identically zero, otherwise it is said that $P(\lambda)$ is *singular*. Matrix polynomials appear in many applications such as mechanics, control theory, computer-aided graphic design and differential algebraic equations. Roughly speaking one can say that regular polynomials appear in mechanics and graphic design [71, 109, 110, 149] and singular polynomials appear essentially in control theory and differential algebraic equations [94, 157, 160].

From a numerical point of view, in the regular case, the main objectives are to compute the eigenvalues and eigenvectors of $P(\lambda)$ [71, 149], and in the singular case, also compute the minimal indices and minimal bases [62, 94, 160]. The *finite eigenvalues* of a regular matrix polynomial are the roots of the scalar polynomial $\det P(\lambda)$, and a vector $v \neq 0$ is an *eigenvector* of $P(\lambda)$ associated to the eigenvalue λ_0 if it satisfies $P(\lambda_0)v = 0$. The definition of eigenvalue in the singular case is more intricate and requires to introduce the concept of *normal rank* of $P(\lambda)$, denoted by $\operatorname{nrank} P(\lambda)$:

$$\operatorname{nrank} P(\lambda) := \max_{\lambda \in \mathbb{C}} \operatorname{rank} P(\lambda).$$

Then we say that λ_0 is a *finite eigenvalue* of $P(\lambda)$ if

$$\operatorname{rank} P(\lambda_0) < \operatorname{nrank} P(\lambda).$$

We say that $P(\lambda)$ has an eigenvalue at ∞ if the *reverse polynomial*

$$\operatorname{rev} P(\lambda) = \lambda^k P(1/\lambda) = \sum_{i=0}^{k} \lambda^i A_{k-i}$$

has $\lambda = 0$ as an eigenvalue. Infinite eigenvalues play an important role in the dynamic behaviour of linear systems that appear in control theory [94, 138]. The concept of eigenvector cannot be extended to singular matrix polynomials. The corresponding concept for pencils is the one of *reducing subspace* [158], whose extension to matrix polynomials is still an open question and is beyond the scope of this work.

The standard way to numerically solve a polynomial eigenvalue problem for a regular matrix polynomial $P(\lambda)$ is to first linearize $P(\lambda)$ into a matrix pencil $L(\lambda) = \lambda X + Y$ with $X, Y \in \mathbf{C}^{nk \times nk}$ [71], and then compute the eigenvalues and eigenvectors of $L(\lambda)$ using well established algorithms for the *generalized eigenvalue problem*, such as the QZ algorithm or some variants of the Arnoldi method [10].

The precise definition of a linearization of a regular polynomial was introduced in [71]. However, a linearization of $P(\lambda)$ does not necessarily have the same elementary divisors at ∞ as $P(\lambda)$. For this reason it was introduced in [72] the concept of strong linearization.

Definition 1.6. *A matrix pencil $L(\lambda) = \lambda X + Y$ with $X, Y \in \mathbb{C}^{nk \times nk}$ is a* linearization *of an $n \times n$ matrix polynomial $P(\lambda)$ of degree k if there exist two unimodular [2] $nk \times nk$ matrices $U(\lambda)$*

[2] A unimodular matrix $V(\lambda)$ is a matrix polynomial such that $\det V(\lambda)$ is a nonzero constant.

and $V(\lambda)$ *such that*

$$U(\lambda)L(\lambda)V(\lambda) = \begin{bmatrix} I_{(k-1)n} & 0 \\ 0 & P(\lambda) \end{bmatrix} \tag{1.33}$$

A linearization $L(\lambda)$ *of* $P(\lambda)$ *is a* strong linearization *if* $\mathrm{rev}L(\lambda)$ *is a linearization of* $\mathrm{rev}P(\lambda)$.

These definitions were introduced in [71, 72] only for regular polynomials, and they were extended in [44] to square singular matrix polynomials. The use of linearizations is justified by the following two facts. First, all linearizations (resp. strong linearizations) of $P(\lambda)$ have the same finite (resp. finite and infinite) elementary divisor [66] as $P(\lambda)$. As was mentioned above, since the linearization procedure transforms a matrix polynomial into a matrix pencil, well-established algorithms for the generalized eigenvalue problem may be used on linearizations both for regular and singular polynomials [42, 43, 55, 73, 159, 160].

The classical approach is to use as linearizations the *first* and *second Frobenius companion forms* [71]. These linearizations are, respectively, $C_1(\lambda) = \lambda X_1 + Y_1$ and $C_2(\lambda) = \lambda X_2 + Y_2$, where $X_1 = X_2 = \mathrm{diag}(A_k, I_n, \ldots, I_n)$ and

$$Y_1 = \begin{bmatrix} A_{k-1} & A_{k-2} & \cdots & A_0 \\ -I_n & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & -I_n & 0 \end{bmatrix}, \qquad Y_2 = \begin{bmatrix} A_{k-1} & -I_n & & 0 \\ A_{k-2} & 0 & \ddots & \\ \vdots & \vdots & \ddots & -I_n \\ A_0 & 0 & \cdots & 0 \end{bmatrix}$$

Notice that $C_1(\lambda)$ and $C_2(\lambda)$ are a natural generalization of the Frobenius companion matrices C_1 and C_2 in (1.2) for scalar polynomials, where the coefficients of the polynomial are now replaced by the matrix coefficients of the matrix polynomial, and the entries identically equal to one are replace by identity blocks.

Although the use of Frobenius companion forms is widely extended in the computation of eigenvalues and eigenvectors of matrix polynomials, this approach presents important disadvantages, such as:

- The loss of the structure. That is, if the matrix polynomial $P(\lambda)$ has some of the structures that arise usually in applications (for example, $P(\lambda)$ is symmetric, skew-symmetric, palindromic, anti-palindromic, even, odd, etc), Frobenius companion forms in general do not share this structure. Therefore the rounding errors inherent to numerical computations may destroy qualitative aspects of the spectrum. One of the best known examples of this phenomenon arises in palindromic problems: a palindromic matrix polynomial has μ as an eigenvalue if and only if $1/\mu$ is also an eigenvalue, a property that may be completely lost when applying the QZ algorithm to one of the Frobenius companion forms. In general, a numerical method that ignores the structure of the matrix polynomial may produce results which are meaningless in physical applications [149].

- Even in the case of matrix polynomials without any special structure, using the Frobenius companion forms in the numerical calculation of the eigenvalues may produce errors, backward and forward errors, much greater than desirable. An ideal algorithm must compute for the matrix polynomial $P(\lambda)$ the eigenvalues that are the exact eigenvalues of a nearby polynomial:

$$\widetilde{P}(\lambda) = \sum_{i=0}^{k} \widetilde{A}_i \lambda^i, \quad \text{with} \quad \|\widetilde{A}_i - A_i\|_2 = O(u)\|A_i\|_2, \quad \text{for } i = 0, 1, \ldots, k, \tag{1.34}$$

  where u denotes the machine epsilon ($u \approx 10^{-16}$ in double precision). These are the expected errors according to the sensitivity of the data and any algorithm that does not satisfy (1.34) is

not considered satisfactory. For example, the QZ algorithm applied directly on the companion forms does not satisfy (1.34) [53, 101].

- If the degree of the polynomial is high, the Frobenius companion forms have a size much greater than the original matrix polynomial, so the computational cost may increase notably.

These drawbacks have motivated in the last few years an intense activity on the development of new classes of linearizations. There are three main sources of new linearizations of matrix polynomials. A first class of linearizations was presented in [105] and further analyzed in [80, 81, 82, 106]. In particular, the conditioning of the eigenvalues of these linearizations and the backward error of polynomial eigenproblems solved by linearizations belonging to this class were studied in [80, 82], while in [81, 106] the construction of structure preserving linearizations is considered.

Other classes of linearizations were introduced and studied in [3, 4], motivated by the use of non-monomial bases for the space of polynomials. Also, in [44] and later in [151] the vector spaces of potential linearizations introduced in [105] are revisited and they are generalized to any degree-graded polynomial bases.

Another source of linearizations of matrix polynomials was introduced in [5, 6]. The linearizations introduced in [5] received the name of *Fiedler linearizations* in [45] since they are a generalization of the Fiedler companion matrices for scalar polynomials introduced in [59]. In [5] it was shown that these linearizations are strong linearizations in the case of a regular polynomial and in [45] it was shown that the Fiedler pencils are still linearizations when the matrix polynomial is a singular square matrix polynomial. In addition, it was shown in [45] that these linearizations can be used to derive the complete eigenstructure of $P(\lambda)$ including the minimal indices and minimal bases in the case of $P(\lambda)$ being singular. The authors showed explicitly how to recover the minimal indices and bases of the matrix polynomial from those of the Fiedler pencils and how to recover the eigenvectors of a regular polynomial from those of these linearizations without any computational cost. The existence of simple recovery procedures for eigenvectors is relevant for deciding whether or not a certain linearization is useful in applications.

The class of Fiedler linearizations does not contain pencils that are symmetric or palindromic when $P(\lambda)$ is symmetric or palindromic, respectively. To overcome this drawback, based on Fiedler pencils, a wider class of linearizations was introduced in [5]. This class does contain linearizations that preserve the symmetric structure [5] and the palindromic structure [46]. These linearizations were named in [30] *generalized Fiedler linearizations* and recovery procedures for eigenvectors, minimal indices and minimal bases for generalized Fiedler linearizations were also presented in [30]. Also, linearizations in this class are strong linearizations for almost every matrix polynomial, and the only matrix polynomials for which not all of these pencil are linearizations are the ones with singular leading and/or zero degree coefficients. Another class of linearizations of matrix polynomials based on Fiedler linearizations was introduced in [164]. In [29], these linearizations were named *Fiedler linearizations with repetition* and recovery formulas for eigenvectors, minimal indices and minimal bases were derived. Several structure preserving linearizations belonging to this new class have been identified, in particular, in [31] linearizations that preserve the palindromic structure are studied, and in [32, 33] linearizations that preserve the symmetric and alternating structures are also studied.

All references on linearizations mentioned so far are restricted to square matrix polynomials. In [47] Fiedler pencils were generalized to rectangular matrix polynomials and this is the first class of new linearizations that has been generalized to rectangular matrix polynomials. In [47] it is proved that Fiedler pencils of rectangular polynomials are always strong linearizations. Moreover, recovery formulas for minimal bases and minimal indices are derived. These formulas are essentially the same ones as for square matrix polynomials, though the techniques used to obtain these formulas are much more involved.

To summarize, there is a recent intense activity towards the development of new classes of linearizations of matrix polynomials. Among all the new families of linearizations that have been

introduced recently, the families of Fiedler linearizations, generalized Fiedler linearizations, and Fiedler linearizations with repetitions possess many algebraic properties that make them particularly interesting and give a strong motivation for studying them:

(a) They are easily constructible from the coefficients of the matrix polynomial.

(b) Fiedler pencils are strong linearizations for every matrix polynomial, and most generalized Fiedler linearizations and Fiedler linearizations with repetition are strong linearizations for every matrix polynomial.

(c) Eigenvectors, minimal indices and minimal bases of the matrix polynomial are easily recovered from those of any of these linearizations.

By contrast, property (b) is not true for the pencils introduced in [4, 105], which are not linearizations for certain regular and singular polynomials. In fact, for odd degree matrix polynomials, the structure preserving generalized Fiedler pencils presented in [5, 46] for symmetric and palindromic polynomials are always strong linearizations, which is again in contrast with the structured pencils developed in [81, 105, 106] that are not linearizations for certain regular polynomials and, in fact, are never linearizations for singular polynomials.

Finally, like in the polynomial-root finding problem using Frobenius companion matrices, the computation of the eigenvalues of a matrix polynomial and the computation of the eigenvalues of one of its linearizations are mathematically equivalent, but they may present different numerical properties, in particular, those regarding conditioning and backward errors. The numerical properties of the linearizations in [105] are very well known [80, 82, 83], but the study of the numerical properties of Fiedler linearizations for matrix polynomials has started very recently [52].

1.6 Organization of the dissertation

This dissertation is organized as follows.

In Chapter 2 we present the definition of Fiedler companion matrices following a different notation to the one used in the original reference [59] in order to better express the results in this work. We also present some basic definitions related with Fiedler matrices that will be used throughout all this dissertation, and some special Fiedler matrices that enjoy several interesting numerical advantages. Finally, we present an algorithm to construct Fiedler matrices, which is a particular case of the algorithm presented in [47], and some basic properties of Fiedler matrices that are immediate consequences of this algorithm. This algorithm will be key in several proofs of the new results presented in this work. Except Proposition 2.20, the results presented in Chapter 2 are not original contributions of the author.

Chapters 3, 4, and 5 are devoted to the study of several algebraic and structural properties of Fiedler matrices. All the results presented in those chapters are original contributions of the author.

In Chapter 3 we study matrix norms of Fiedler matrices and their inverses. We obtain explicit expressions for the 1-,∞- and Frobenius norms of any Fiedler matrix and its inverse. These norms are obtained through the algorithm presented in Chapter 2 to construct Fiedler matrices, together with a new algorithm presented in this chapter to construct inverses of Fiedler matrices. These norms will be used in Chapter 6 to obtain new and, for a wide family of polynomials, sharper than previous ones lower and upper bounds for the absolute value of the roots of a polynomial. The results in this chapter have been published in [50].

The study of the spectral norm, or 2-norm, of Fiedler matrices and their inverses is postponed to Chapter 4. The spectral norms of a matrix A and its inverse A^{-1} are, respectively, the largest and the reciprocal of the smallest singular values of A. For this reason, Chapter 4 is devoted to

the study of singular values of Fiedler matrices. More precisely, we determine how many of their singular values are equal to 1 and, for those that are not, we show that they can be obtained from the square roots of the eigenvalues of certain matrices that may have sizes much smaller than $n \times n$ and that are easily constructible from the coefficients of $p(z)$. This result is the main contribution of this chapter. The are two key tools to prove this result. The first one is the concept of *staircase matrix*, introduced in Section 4.1. Staircase matrices are matrices whose nonzero entries follow a very special pattern. This concept, as far as we know, is new in the literature. The second tool is the result stating that any Fiedler matrix can be expressed as a sum of a permutation matrix plus a matrix whose rank varies from 1 to $\lfloor (n+1)/2 \rfloor$. In addition, we show how to construct these two summands via simple algorithms and how to determine the rank of the second summand based on the properties of staircase matrices. In plain words, we show that any Fiedler matrix admit expressions as "unitary plus low-rank matrices". The results in this chapter have been published in [49].

In Chapter 5 we present two expressions for the adjugate matrix of $zI - M_\sigma$, where M_σ is a Fiedler matrix. These formulas are original contributions obtained by the author except in the case when M_σ is one of the Frobenius companion matrices, which were previously obtained in [150]. These expressions for the adjugate matrix of $zI - M_\sigma$ will be one of the key tools used in Chapters 8 and 9, where we study eigenvalue condition numbers and pseudospectra of Fiedler matrices, and the backward errors of computed roots of monic polynomials using Fiedler matrices. The results in this chapter have been published in [51].

Chapter 6 is devoted to the applications of Fiedler matrices in finding bounds for the roots of monic polynomials. We investigate the lower and upper bounds for the absolute values of the roots of monic polynomials that are obtained from the norms of Fiedler matrices presented in Chapter 3. We show that norms of Fiedler matrices produce many new bounds, but none of them improve significantly the classical bounds obtained from the Frobenius companion matrices described in Section 1.4. However, we prove that if the norms of the inverses of Fiedler matrices are used, then another family of new bounds is obtained and some of the bounds in this family improve significantly the bounds coming from the Frobenius companion matrices for certain polynomials. We also present some theoretical results concerning the best bounds that may be obtained from diagonal similarities of Fiedler matrices. The results of this chapter have been published in [50].

Chapters 7, 8, and 9 are devoted to the study of the numerical properties that are interesting for applying Fiedler matrices in numerical computations (like numerical methods for computing roots of polynomials):

(a) condition numbers for inversion,

(b) eigenvalue condition numbers and pseudospectra of Fiedler matrices, and

(c) backward errors of the computed roots of a polynomial using Fiedler companion matrices.

All the results presented in those chapters are original contributions by the author.

In Chapter 7 we investigate condition numbers for inversion of Fiedler matrices. More precisely, we present explicit expressions for the condition numbers for inversion of all Fiedler matrices with respect to the Frobenius norm. This allows us to get a very simple criterion for ordering all Fiedler matrices according to increasing condition numbers and to provide lower and upper bounds on the ratio of the condition numbers of any pair of Fiedler matrices. These results establish that if $|p(0)| \leq 1$, then the Frobenius companion matrices have the largest condition number among all Fiedler matrices of $p(z)$, and that if $|p(0)| > 1$, then the Frobenius companion matrices have the smallest condition number. We also provide families of polynomials where the ratio of the condition numbers of pairs of Fiedler matrices can be arbitrarily large and prove that this can only happen when both Fiedler matrices are very ill-conditioned. The results in this chapter has been published in [49].

In Chapter 8 we investigate eigenvalue condition numbers and pseudospectra of Fiedler matrices of a monic polynomial $p(z)$. We present explicit expressions for the eigenvalue condition numbers for any Fiedler matrix, and then we compare these condition numbers with the condition numbers of the roots of the original polynomial. We show that if the maximum of the absolute values of the coefficients of $p(z)$ is much larger or much smaller than 1, then the eigenvalues of Fiedler matrices may be potentially much more ill conditioned than the roots of $p(z)$. By contrast, if the maximum of the absolute values of the coefficients of $p(z)$ is moderate and not close to zero, that is, it is of order $\Theta(1)$, then the eigenvalues of Fiedler matrices and the roots of $p(z)$ are guaranteed to have similar condition numbers, and therefore, from the point of view of eigenvalue condition numbers, in this case all Fiedler companion matrices are good tools for the purpose of computing roots of monic polynomials. We then study the ratio between eigenvalue condition numbers of Frobenius companion matrices and eigenvalue condition numbers of Fiedler matrices other than the Frobenius ones. We show that if the absolute value of the coefficients of $p(z)$ are moderate then this ratio is also moderate and, therefore, from the point of view of condition numbers, in this situation any Fiedler matrix can be used for solving the root-finding problem for $p(z)$ with the same reliability as Frobenius companion matrices. By contrast, this ratio may be potentially large or small for polynomials with large coefficients. In this case, from the point of view of condition numbers, for some polynomials with large coefficients some Fiedler matrices may be more convenient than the Frobenius ones and, on the contrary, for some other polynomials with large coefficients, Frobenius companion matrices may be more convenient than other Fiedler matrices. However, we show that, from the point of view of condition numbers, Frobenius companion matrices are better suited than the rest of Fiedler matrices in the problem of computing roots of monic polynomial only for polynomials for which it is not recommended to compute their roots as eigenvalues of any Fiedler matrices (including the Frobenius ones). Finally, we show that there are polynomials for which one should avoid computing their roots as the eigenvalues of Frobenius companion matrices and to use, instead, another Fiedler matrix. Although how to identify these polynomials and how to know which Fiedler matrix one might use instead of the Frobenius ones are still an ongoing works.

Regarding pseudospectra of Fiedler matrices, first, we show how to accurate estimate them in a $m \times m$ grid using only $O(nm^2)$ flops compared with the $O(n^3m^2)$ flops needed in the SVD method explained in Section 1.2.2.2. Then, we establish various mathematical relationships between the pseudozero sets of a monic polynomial $p(z)$ and the pseudospectra of the associated Fiedler matrices.

Finally, the effect of balancing Fiedler matrices is also investigated from the point of view of eigenvalue condition numbers and pseudospectra. We present numerical evidence that shows the following: if Fiedler matrices are balanced then the roots of $p(z)$ and the eigenvalues of the balanced Fiedler matrices are usually equally conditioned, and that pseudozero sets of $p(z)$ and pseudospectra of Fiedler matrices are usually quite close to each other. We want to emphasize that the results in Chapter 8 have not been published yet and that we are preparing a paper that we expect to submit soon.

In Chapter 9 we analyze the backward stability of polynomial root-finding algorithms via Fiedler companion matrices. In other words, given a monic polynomial $p(z)$, the question is to determine whether the whole set of computed eigenvalues of a Fiedler companion matrix, computed with a backward stable algorithm for the standard eigenvalue problem, is the set of roots of a nearby polynomial or not. We show that, if the coefficients of $p(z)$ are bounded in absolute value by a moderate number, then algorithms for polynomial root-finding using Fiedler matrices are backward stable from the polynomial point of view, and Fiedler matrices are as good as the Frobenius companion matrices for the purpose of computing roots of monic polynomials. This would allow us to use Fiedler companion matrices with favorable structures in the polynomial root-finding problem. However, when some of the coefficients of the polynomial is very large, companion Fiedler matrices may produce larger backward errors than Frobenius companion matrices, although in this case neither Frobenius nor Fiedler matrices lead to backward stable computations. To prove

this we obtain explicit expressions for the change, to first order, of the characteristic polynomial coefficients of Fielder matrices under small perturbations. Also, the effect of balancing Fiedler matrices is studied. We present numerical evidence that shows that, if Fiedler matrices are balanced, computing the roots of a monic polynomial via the eigenvalues of Fiedler matrices and a backward stable eigensolver is usually a backward stable method from the point of view of polynomials. The results in this chapter has been published in [51].

Chapter 2

Definition and basic properties of Fiedler matrices

This chapter is devoted to the definition of *Fiedler matrices*, to establish their basic properties, and introduce some definitions related with them. The properties of Fiedler matrices stated in this chapter are not original contributions of the author, except Proposition 2.20. We also present some particular Fiedler matrices that will appear throughout all this work.

2.1 Definition of Fiedler matrices

For a given monic polynomial $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ with $a_i \in \mathbb{C}$ Fiedler matrices are constructed in [59] as the product

$$M_{i_1} M_{i_2} \cdots M_{i_n},$$

where the matrices $M_0, M_1, \ldots, M_{n-1}$ are defined in (1.3), and $(i_1, i_2, \ldots, i_n)$ is any possible permutation of the n-tuple $(0, 1, \ldots, n-1)$. In order to better express certain key properties of this permutation and the resulting Fiedler matrix, in [45] the authors index the product of the M_i factors in a slightly different way, as it is described in the following definition.

Definition 2.1. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$, with $n \geq 2$, and let M_i, for $i = 0, 1, \ldots, n-1$, be the matrices defined in* (1.3)*. Given any bijection $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$, the* Fiedler matrix of $p(z)$ associated with σ *is the $n \times n$ matrix*

$$M_\sigma := M_{\sigma^{-1}(1)} \cdots M_{\sigma^{-1}(n)}. \tag{2.1}$$

Remark 2.2. Sometimes we will write the bijection σ using the array notation $\sigma = (\sigma(0), \sigma(1), \ldots, \sigma(n-1))$.

We want to notice that $\sigma(i)$ in (2.1) describes the position of the factor M_i in the product $M_{\sigma^{-1}(1)} \cdots M_{\sigma^{-1}(n)}$, i.e., $\sigma(i) = j$ means that M_i is the jth factor in the product. We want to note also that the building factors (1.3) of (2.1) depend on $p(z)$ (to be precise, they depend on its coefficients). However, in this case we skip this dependence for the sake of simplicity. Finally, when necessarily, we will indicate explicitly the dependence of M_σ on a certain polynomial $q(z)$ using the notation: $M_\sigma(q)$.

Example 2.3. Let $p(z) = z^5 + \sum_{k=0}^{4} a_k z^k$ be a monic polynomial of degree 5. Then the following

matrices are Fiedler matrices of $p(z)$:

$$M_{\sigma_1} = M_0M_1M_2M_3M_4 = \begin{bmatrix} -a_4 & 1 & 0 & 0 & 0 \\ -a_3 & 0 & 1 & 0 & 0 \\ -a_2 & 0 & 0 & 1 & 0 \\ -a_1 & 0 & 0 & 0 & 1 \\ -a_0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$M_{\sigma_2} = M_4M_2M_0M_1M_3 = \begin{bmatrix} -a_4 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -a_2 & 0 & -a_1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_0 & 0 \end{bmatrix} \quad \text{and}$$

$$M_{\sigma_3} = M_2M_1M_0M_3M_4 = \begin{bmatrix} -a_4 & 1 & 0 & 0 & 0 \\ -a_3 & 0 & 1 & 0 & 0 \\ -a_2 & 0 & 0 & -a_1 & -a_0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Also, according to the notation explained in Remark 2.2, we have $\sigma_1 = (1, 2, 3, 4, 5)$, $\sigma_2 = (3, 4, 2, 5, 1)$ and $\sigma_3 = (3, 2, 1, 4, 5)$.

The family of matrices $\{M_k\}_{k=0}^{n-1}$ in (1.3) satisfies the following commutativity relations

$$M_iM_j = M_jM_i \quad \text{for } |i - j| \neq 1, \tag{2.2}$$

that can be easily checked. The relations (2.2) imply that some Fiedler matrices associated with different bijections σ are equal. For example, for $n = 3$, the Fiedler matrices $M_0M_2M_1$ and $M_2M_0M_1$ are equal. These relations suggest that the relative positions of the matrices M_i and M_{i+1} in the product M_σ are of fundamental interest in studying Fiedler matrices. This motivates Definition 2.4, introduced in [45].

Definition 2.4. *Let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection.*

(a) *For $i = 0, \ldots, n-2$, we say that σ has a* consecution *at i if $\sigma(i) < \sigma(i+1)$ and that σ has an* inversion *at i if $\sigma(i) > \sigma(i+1)$.*

(b) *The* positional consecution-inversion sequence of σ, *denoted by* $\mathrm{PCIS}(\sigma)$, *is the $(n-1)$-tuple $(v_0, v_1, \ldots, v_{n-2})$ such that $v_j = 1$ if σ has a consecution at j and $v_j = 0$ if σ has an inversion at j.*

(c) *The* consecution-inversion structure sequence of σ, *denoted by* $\mathrm{CISS}(\sigma)$, *is the tuple $(\mathfrak{c}_0, \mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{c}_\ell, \mathfrak{i}_\ell)$, where σ has $\mathfrak{c}_0$ consecutive consecutions at $0, 1, \ldots, \mathfrak{c}_0 - 1$; $\mathfrak{i}_0$ consecutive inversions at $\mathfrak{c}_0, \mathfrak{c}_0 + 1, \ldots, \mathfrak{c}_0 + \mathfrak{i}_0 - 1$ and so on, up to $\mathfrak{i}_\ell$ inversions at $n - 1 - \mathfrak{i}_\ell, \ldots, n - 2$.*

Remark 2.5. We will use the following simple observation on the concepts introduced in Definition 2.4 without explicitly referring to.

1. Part (a) is related to the matrix M_σ as follows: σ has a consecution at i if and only if M_i is to the left of M_{i+1} in the product defining the Fiedler matrix M_σ, while σ has an inversion at i if and only if M_i is to the right of M_{i+1} in M_σ.

2. Note that $\mathfrak{c}_0$ and $\mathfrak{i}_\ell$ in $\mathrm{CISS}(\sigma)$ may be zero (in the first case, σ has an inversion at 0 and in the second one it has a consecution at $n-2$) but $\mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{i}_{\ell-1}, \mathfrak{c}_\ell$ are all strictly positive. These conditions uniquely determine $\mathrm{CISS}(\sigma)$ and, in particular, the parameter ℓ.

The tuples PCIS(σ), and CISS(σ) are equivalent, i.e., if any of them is given then the other one can be easily obtained. Also, the two tuples are uniquely determined by σ, but none of them determine uniquely σ, in general.

Example 2.6. Consider the Fiedler matrix $M_\sigma = M_6M_5M_3M_0M_1M_2M_4M_7M_8$ of the monic degree-9 polynomial $p(z) = z^9 + \sum_{k=0}^{8} a_k z^k$. Then, σ is a bijection such that PCIS(σ) $= (1,1,0,1,0,0,1,1)$ and CISS(σ) $= (2,1,1,2,2,0)$.

Due to the commutativity relations (2.2), the only needed information to construct the Fiedler matrix M_σ is the polynomial $p(z)$ and PCIS(σ) or, equivalently, the polynomial $p(z)$ and CISS(σ). On the opposite way, given a polynomial $p(z)$ one may think that if two Fiedler companion matrices M_{σ_1} and M_{σ_2} of $p(z)$ are equal then PCIS(σ_1) = PCIS(σ_2), but the fact that $M_0 = I$ when $a_0 = -1$ complicates a result in this line.

Proposition 2.7. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with $n \geq 2$, let $\sigma_1, \sigma_2 : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be two bijections, and let M_{σ_1} and M_{σ_2} be the Fiedler matrices of $p(z)$ associated with the bijections σ_1 and σ_2, respectively.*

(a) *Suppose $a_0 \neq -1$. Then, $M_{\sigma_1} = M_{\sigma_2}$ if and only if* PCIS(σ_1) = PCIS(σ_2).

(b) *Suppose $a_0 = -1$. Then, $M_{\sigma_1} = M_{\sigma_2}$ if and only if the last $n-2$ entries of* PCIS(σ_1) *are equal to the last $n-2$ entries of* PCIS(σ_2).

For quick references we include here other basic definitions related with PCIS(σ) that we will use in future chapters.

Definition 2.8. *Let $\sigma : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be a bijection with* PCIS(σ) $= (v_0, v_1, \ldots, v_{n-2})$ *and with* CISS(σ) $= (\mathfrak{c}_0, \mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{c}_\ell, \mathfrak{i}_\ell)$, *then:*

(a) *The* number of initial consecutions or inversions of σ, *denoted by t_σ, is*

$$t_\sigma = \begin{cases} \mathfrak{c}_0 & \text{if } \mathfrak{c}_0 \neq 0, \\ \mathfrak{i}_0 & \text{if } \mathfrak{c}_0 = 0. \end{cases}$$

(b) *The* reduced consecution-inversion structure sequence of σ, *denoted by* RCISS(σ), *is the sequence obtained from* CISS(σ) *after removing the zero entries.*

(c) *The* extended positional consecution-inversion sequence of σ, *denoted by* EPCIS(σ), *is the n-tuple $(v_0, v_1, \ldots, v_{n-1})$, where $v_{n-1} = v_{n-2}$.*

(d) *For $0 \leq i \leq j \leq n-2$, we set*

$$\mathfrak{i}_\sigma(i:j) := \sum_{k=i}^{j}(1 - v_k) \qquad \text{and} \qquad \mathfrak{c}_\sigma(i:j) := \sum_{k=i}^{j} v_k$$

for, respectively, the number of inversions and consecutions of σ from i to j. We also set $\mathfrak{i}_\sigma(i:j) := 0$ and $\mathfrak{c}_\sigma(i:j) := 0$ for $i > j$.

Remark 2.9. The following simple observations on Definition 2.8 will be used.

1. The functions $\mathfrak{i}_\sigma(i:j)$ and $\mathfrak{c}_\sigma(i:j)$ satisfy the following identities:

$$\mathfrak{i}_\sigma(i:j) + \mathfrak{c}_\sigma(i:j) = j - i + 1, \qquad \text{for } 0 \leq i \leq j \leq n-2, \tag{2.3}$$

$$\mathfrak{i}_\sigma(0:i) + \mathfrak{c}_\sigma(0:j) \leq n-1, \qquad \text{for } 0 \leq i, j \leq n-2. \tag{2.4}$$

2. According to the comment 2 in Remark 2.5, $\mathrm{RCISS}(\sigma) = \mathrm{CISS}(\sigma)$ if and only if $\mathfrak{c}_0 \neq 0$ and $\mathfrak{i}_\ell \neq 0$.

3. $1 \leq t_\sigma \leq n-1$

Example 2.10. Consider the Fiedler matrix M_σ in Example 2.6. Then, σ is a bijection such that $t_\sigma = 2$, $\mathrm{EPCIS}(\sigma) = (1,1,0,1,0,0,1,1,1)$ and $\mathrm{RCISS}(\sigma) = (2,1,1,2,2)$.

In Chapter 6 we will use the concept of *reversal bijection*.

Definition 2.11. *Given a bijection $\sigma : \{0,1,\ldots,n-1\} \longrightarrow \{1,\ldots,n\}$, the* reversal bijection *of σ, denoted by* $\mathrm{rev}(\sigma) : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$, *is defined by* $\mathrm{rev}(\sigma)(i) = n+1-\sigma(i)$, *for* $0 \leq i \leq n-1$.

Since the matrix M_k, for $k = 0,1,\ldots,n-1$, in (1.3) is a symmetric matrix, it follows that the transpose of a Fiedler matrix is also a Fiedler matrix. The following result, whose easy proof is omitted, relates the transpose of a Fiedler matrix M_σ with the reversal bijection of σ.

Theorem 2.12. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$, with $n \geq 2$, let $\sigma : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be a bijection, let* $\mathrm{rev}(\sigma)$ *be the reversal bijection of σ, and let M_σ and $M_{\mathrm{rev}(\sigma)}$ be the Fiedler matrices of $p(z)$ associated with σ and* $\mathrm{rev}(\sigma)$, *respectively. Then,*

$$M_\sigma^T = M_{\mathrm{rev}(\sigma)}. \tag{2.5}$$

We close this section with the following notions, not strictly related to Fiedler matrices, that will be used along several chapters.

Definition 2.13. *Let $p(z) = \sum_{k=0}^{n} a_k z^k$ be a polynomial of degree n. For $d = 0,1,\ldots,n$, the* degree d Horner shift of $p(z)$ *is the polynomial $p_d(z) := a_n z^d + a_{n-1} z^{d-1} + \cdots + a_{n-d+1} z + a_{n-d}$.*

Notice that $p_n(z) = p(z)$ and that we do not assume that $p(z)$ is monic. Also, the Horner shifts of $p(z)$ satisfy the following recurrence relation:

$$\begin{cases} p_0(z) = a_n\,, \quad \text{and} \\ p_d(z) = z p_{d-1}(z) + a_{n-d}\,, \quad \text{for } \ d = 1,2,\ldots,n-1. \end{cases} \tag{2.6}$$

Definition 2.14. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n. If $a_0 \neq 0$, then the* monic reversal polynomial *of $p(z)$, denoted by $p^\sharp(z)$, is defined as:*

$$p^\sharp(z) := \frac{z^n}{a_0} p(z^{-1}) = z^n + \frac{a_1}{a_0} z^{n-1} + \frac{a_2}{a_0} z^{n-2} + \cdots + \frac{a_{n-1}}{a_0} z + \frac{1}{a_0},$$

and the reversal polynomial *of $p(z)$, denoted by $p^{\mathrm{rev}}(z)$, is defined as:*

$$p^{\mathrm{rev}}(z) = z^n p(z^{-1}) = a_0 z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_{n-1} z + 1.$$

Observe that the roots of $p^\sharp(z)$ and $p^{\mathrm{rev}}(z)$ are the reciprocals of the roots of $p(z)$.

Definition 2.15. *Given a monic polynomial $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$, and a nonzero complex number α, we define the monic scaled polynomial of $p(z)$, denoted by $p_\alpha(z)$, as*

$$p_\alpha(z) = \frac{1}{\alpha^n} p(\alpha z) = z^n + \sum_{k=0}^{n-1} \frac{a_k}{\alpha^{n-k}} z^k.$$

Note that the roots of $p(z)$ may be recovered from the roots of $p_\alpha(z)$, that is, if λ is a root of $p_\alpha(z)$ then $\alpha\lambda$ is a root of $p(z)$.

2.2 Relevant examples of Fiedler matrices

Given a monic polynomial $p(z)$ as in (1.1), the set of Fiedler matrices of $p(z)$ includes the first and second Frobenius companion matrices of $p(z)$. That is, $C_1 = M_{n-1}M_{n-2}\cdots M_1M_0$ and $C_2 = M_0M_1\cdots M_{n-2}M_{n-1}$. The first Frobenius companion matrix C_1 is a Fiedler matrix associated with the bijection $\sigma_1 = (n, n-1, \ldots, 2, 1)$. For this bijection we have that $\mathrm{PCIS}(\sigma_1) = (0, 0, \ldots, 0)$, $\mathrm{CISS}(\sigma_1) = (0, n-1)$, $\mathrm{RCISS}(\sigma_1) = (n-1)$ and $t_{\sigma_1} = n-1$. The second Frobenius companion matrix C_2 is a Fiedler matrix associated with the bijection $\sigma_2 = (1, 2, \ldots, n-1, n)$. For this bijection we have that $\mathrm{PCIS}(\sigma_2) = (1, 1, \ldots, 1)$, $\mathrm{CISS}(\sigma_2) = (n-1, 0)$, $\mathrm{RCISS}(\sigma_2) = (n-1)$ and $t_{\sigma_2} = n-1$.

For any value of the degree of a monic polynomial, the set of Fiedler matrices also includes four pentadiagonal matrices that have a much smaller bandwidth that the first and second Frobenius companion matrices (if the degree n of the polynomial is high). These four pentadiagonal Fiedler matrices are constructed as follows: let $B = M_1M_3\cdots$ be the product of the odd M_i factors and let $C = M_2M_4\cdots$ be the product of the even M_i factors with the exception of M_0. Then, is easy to check that the product of M_0, B and C in any order yields a pentadiagonal Fiedler matrix. Since M_0 and C commute we have only four different matrices, namely,

$$P_1 := M_0CB, \quad P_2 := CBM_0, \quad P_3 := BM_0C \quad \text{and} \quad P_4 := M_0BC. \tag{2.7}$$

Using the commutativity relations in (2.2) and Theorem 2.12, is easy to check that $P_1^T = P_3$ and $P_2^T = P_4$. If we consider a degree-6 monic polynomial $p(z) = z^6 + \sum_{k=0}^{5} a_k z^k$, P_1 and P_2 are

$$P_1 = \begin{bmatrix} -a_5 & 1 & 0 & 0 & 0 & 0 \\ -a_4 & 0 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_2 & 0 & -a_1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -a_0 & 0 \end{bmatrix} \quad \text{and} \quad P_2 = \begin{bmatrix} -a_5 & 1 & 0 & 0 & 0 & 0 \\ -a_4 & 0 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_2 & 0 & -a_1 & -a_0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

and if we consider a degree-7 monic polynomial $p(z) = z^7 + \sum_{k=0}^{6} a_k z^k$, then P_1 and P_2 are

$$P_1 = \begin{bmatrix} -a_6 & -a_5 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -a_4 & 0 & -a_3 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_2 & 0 & -a_1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -a_0 & 0 \end{bmatrix} \quad \text{and} \quad P_2 = \begin{bmatrix} -a_6 & -a_5 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -a_4 & 0 & -a_3 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_2 & 0 & -a_1 & -a_0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

In general, the matrix P_1 in (2.7) is a Fiedler matrix associated with a bijection μ_1 such that

$$\mathrm{PCIS}(\mu_1) = \begin{cases} (1, 0, 1, 0, \ldots, 1, 0) & \text{if} \quad n \text{ is odd}, \\ (1, 0, 1, 0, \ldots, 1, 0, 1) & \text{if} \quad n \text{ is even}, \end{cases}$$

or equivalently,

$$\mathrm{CISS}(\mu_1) = \begin{cases} (1, 1, \ldots, 1, 1) \in \mathbb{R}^{n-1} & \text{if} \quad n \text{ is odd}, \\ (1, 1, \ldots, 1, 0) \in \mathbb{R}^{n} & \text{if} \quad n \text{ is even}, \end{cases}$$

Also, for this bijection we have that $\mathrm{RCISS}(\mu_1) = (1, 1, \ldots, 1) \in \mathbb{R}^{n-1}$ and $t_{\mu_1} = 1$. The matrix P_2 in (2.7) is a Fiedler matrix is associated with a bijection μ_2 such that

$$\mathrm{PCIS}(\mu_2) = \begin{cases} (0, 0, 1, 0, \ldots, 1, 0) & \text{if} \quad n \text{ is odd}, \\ (0, 0, 1, 0, \ldots, 1, 0, 1) & \text{if} \quad n \text{ is even}, \end{cases}$$

or equivalently,

$$\mathrm{CISS}(\mu_2) = \begin{cases} (0,2,1,1,\ldots,1,1) \in \mathbb{R}^{n-1} & \text{if} \quad n \text{ is odd,} \\ (0,2,1,1,\ldots,1,0) \in \mathbb{R}^{n} & \text{if} \quad n \text{ is even,} \end{cases}$$

Also, for this bijection μ_2 we have that $\mathrm{RCISS}(\mu_2) = (2,1,1,\ldots,1) \in \mathbb{R}^{n-2}$ and $t_{\mu_2} = 2$.

Excluding the Frobenius companion matrices, the simplest Fiedler matrices are those corresponding to bijections with just one inversion (resp., consecution) at 0, and consecutions (resp., inversions) elsewhere. These particular Fiedler matrices present several numerical advantages that may be of interest in new enhancements of the current codes for the Polynomial Eigenvalue Problem (like MATLAB's `polyeig`). To be precise, one of these matrices is

$$F := M_1 M_2 \cdots M_{n-1} M_0 = \begin{bmatrix} -a_{n-1} & 1 & & & \\ \vdots & & \ddots & & \\ -a_2 & & & 1 & \\ -a_1 & & & & -a_0 \\ 1 & & & & \end{bmatrix}. \tag{2.8}$$

and the other one is F^T. This Fiedler matrix will play a relevant role in future sections (specially in Chapters 6 and 7). The matrix F is a Fiedler matrix associated with a bijection τ such that $\mathrm{PCIS}(\tau) = (0,1,1,\ldots,1)$, $\mathrm{CISS}(\tau) = (0,1,n-2,0)$, $\mathrm{RCISS}(\tau) = (1,n-2)$ and $t_\tau = 1$.

2.3 A multiplication free algorithm to construct Fiedler matrices and its consequences

To construct a Fiedler matrix M_σ, the obvious way is to perform the multiplication of all the M_i factors directly, but in [47, Algorithm 1], the authors give an algorithm which constructs Fiedler matrices without performing any arithmetic operation. `Algorithm 1` in [47] considers the general case of Fiedler linearizations of matrix polynomials, not necessarily monic. In Theorem 2.16 we recall this algorithm only for monic scalar polynomials. Here, and in the rest of this work, we use MATLAB *notation for submatrices*, that is, $A(i:j,:)$ indicates the submatrix of A consisting of rows i through j and $A(:,k:l)$ indicates the submatrix of A consisting of columns k through l.

Theorem 2.16. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ with $n \geq 2$, let $\sigma : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. Then* `Algorithm 1` *constructs M_σ.*

`Algorithm 1`. *Given $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ and a bijection σ, the following algorithm constructs M_σ.*

if σ has a consecution at 0 then

$$W_0 = \begin{bmatrix} -a_1 & 1 \\ -a_0 & 0 \end{bmatrix}$$

else

$$W_0 = \begin{bmatrix} -a_1 & -a_0 \\ 1 & 0 \end{bmatrix}$$

endif

for $i = 1 : n-2$

if σ has a consecution at i then

$$W_i = \begin{bmatrix} -a_{i+1} & 1 & 0 \\ W_{i-1}(:,1) & 0 & W_{i-1}(:,2:i+1) \end{bmatrix}$$

else

$$W_i = \begin{bmatrix} -a_{i+1} & W_{i-1}(1,:) \\ 1 & 0 \\ 0 & W_{i-1}(2:i+1,:) \end{bmatrix}$$

endif
endfor
$M_\sigma = W_{n-2}$

Remark 2.17. Note that, for $i = 1, 2, \ldots, n-1$, the matrix W_i in `Algorithm 1` is a Fiedler matrix of the polynomial $z^{i+2} + \sum_{k=0}^{i+1} a_k z^k$ associated with a bijection $\rho = (\sigma(0), \sigma(1), \ldots, \sigma(i+1))$.

The interest of this algorithm, apart from constructing Fiedler matrices without performing any arithmetic operation, is that it allows us to prove easily some elementary properties of Fiedler matrices. For instance, since `Algorithm 1` performs $n-1$ "if" decisions, there are at most 2^{n-1} different Fiedler matrices associated with any $p(z)$ of degree $n \geq 2$. In fact, with a little bit of extra effort, the reader may prove, by induction on W_i, that if $a_0 \neq -1$, then all these 2^{n-1} Fiedler matrices are really different, i.e., different for any set of specific values of the coefficients $a_0, a_1, \ldots, a_{n-1}$. However, if $a_0 = -1$, then `Algorithm 1` produces the same W_0 for σ having either a consecution or an inversion at 0, and there are only 2^{n-2} different Fiedler matrices. We summarize these results without proof in Corollary 2.18.

Corollary 2.18. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ with $n \geq 2$.*

(a) *If $a_0 \neq -1$, then there are 2^{n-1} different Fiedler matrices associated with $p(z)$.*

(b) *If $a_0 = -1$, then there are 2^{n-2} different Fiedler matrices associated with $p(z)$.*

For example, for $n = 3$, that is, for cubic polynomials, there are four Fiedler matrices, namely,

$$\begin{bmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} -a_2 & 1 & 0 \\ -a_1 & 0 & 1 \\ -a_0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} -a_2 & -a_1 & 1 \\ 1 & 0 & 0 \\ 0 & -a_0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -a_2 & 1 & 0 \\ -a_1 & 0 & -a_0 \\ 1 & 0 & 0 \end{bmatrix}.$$

From `Algorithm 1`, it is also very easy to prove Theorem 2.19 via an straightforward induction on the matrices W_i. We omit the proof, since Theorem 2.19 is a particular case of the much more general result [47, Theorem 3.10], which proves several structural properties of Fiedler linearizations of rectangular matrix polynomials.

Theorem 2.19. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. Then:*

(a) *M_σ has n entries equal to $-a_0, -a_1, \ldots, -a_{n-1}$, with exactly one copy of each.*

(b) *M_σ has $n-1$ entries equal to 1.*

(c) *The rest of the entries of M_σ are equal to 0.*

(d) *If an entry equal to 1 of those in part* (b) *is at position (i, j), then either the rest of the entries in the ith row of M_σ are equal to 0 or the rest of the entries in the jth column of M_σ are equal to 0.*

(e) *M_σ has either a row (if σ has a consecution at 0) or a column (if σ has an inversion at 0) whose entries are $-a_0$ together with $n-1$ zeros.*

In particular, Theorem 2.19 establishes the fact that any Fiedler matrix has the same entries as the first and second Frobenius companion forms, although placed on different positions.

Another important basic property of Fiedler matrices of a monic polynomial $p(z)$ is that they are *irreducible* matrices[1], when $p(0) \neq 0$. We use in Proposition 2.20 the concept of *directed graph of a matrix* as defined in [87, Definition 6.2.11].

Proposition 2.20. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. Then,*

(a) *The directed graph of M_σ has a cycle that visits all nodes if and only if $a_0 \neq 0$.*

(b) *M_σ is an irreducible matrix if and only if $a_0 \neq 0$.*

Proof. In the proof we will use Part (e) in Theorem 2.19. We will denote by $\Gamma(A)$ the directed graph of a matrix A.

Proof of part (a). If $a_0 = 0$, then M_σ has either a row or a column with all its entries equal to zero. In both cases the corresponding vertex cannot be visited by a cycle and, therefore, $\Gamma(M_\sigma)$ has not a cycle visiting all nodes.

If $a_0 \neq 0$, then we proceed by induction on the matrices W_i defined in `Algorithm 1`. The result is obviously true for W_0 since the entries $W_0(1,2)$ and $W_0(2,1)$ are both different from zero and, so, $\Gamma(W_0)$ has a cycle visiting all nodes. Let us assume that the result is true for the $(i+1)\times(i+1)$ matrix W_{i-1}. We need to distinguish two cases: σ has a consecution at i or σ has an inversion at i. We only prove the result in the case when σ has a consecution at i, since the other one is similar. The fact that $\Gamma(W_{i-1})$ has a cycle that visits all nodes is equivalent to the fact that there exists a permutation $(j_2, j_3, \ldots, j_{i+1})$ of the indices $(2, 3, \ldots, i+1)$ such that

$$W_{i-1}(1, j_2)\, W_{i-1}(j_2, j_3)\, W_{i-1}(j_3, j_4) \cdots W_{i-1}(j_i, j_{i+1})\, W_{i-1}(j_{i+1}, 1) \neq 0. \tag{2.9}$$

The expression of W_i in terms of W_{i-1} given in `Algorithm 1` allows us to write (2.9) in terms of entries of W_i as follows

$$W_i(2, j_2+1)\, W_i(j_2+1, j_3+1)\, W_i(j_3+1, j_4+1) \cdots W_i(j_i+1, j_{i+1}+1)\, W_i(j_{i+1}+1, 1) \neq 0$$

and, since $W_i(1,2) = 1$, we get

$$W_i(1,2)\, W_i(2, j_2+1)\, W_i(j_2+1, j_3+1)\, W_i(j_3+1, j_4+1) \cdots W_i(j_i+1, j_{i+1}+1)\, W_i(j_{i+1}+1, 1) \neq 0,$$

which corresponds to a cycle that visits all nodes in $\Gamma(W_i)$.

Proof of part (b). If $a_0 = 0$, then M_σ has either a row or a column with all its entries equal to zero. If M_σ has a zero row, then select a permutation matrix Π such that $\Pi^T M_\sigma(p)\Pi$ has the nth row equal to zero and we see by definition that M_σ is reducible. If M_σ has a zero column, then select a permutation matrix Π such that $\Pi^T M_\sigma \Pi$ has the nth column equal to zero and we see by definition that M_σ is reducible.

If $a_0 \neq 0$, then, by part (a), $\Gamma(M_\sigma)$ is strongly connected [87, Definition 6.2.13] and this equivalent to the fact that M_σ is irreducible [87, Theorem 6.2.24]. □

[1] We recall that a matrix is irreducible when is not permutation-similar to a block upper triangular matrix.

Chapter 3

Inverses and norms of Fiedler matrices

The goal of this chapter is to obtain explicit expressions for some relevant matrix norms of Fiedler matrices and their inverses. We obtain these expressions in the case of the 1-, ∞- and Frobenius norms in Theorem 3.9, Theorem 3.8 and Corollary 3.5, respectively. We leave the study of the 2-norm to Chapter 4, which is devoted to the study of the singular values of Fiedler matrices. These explicit expressions of norms of Fiedler matrices, which are interesting by themselves, will be used in Chapter 6 to obtain new lower and upper bounds for roots of monic polynomials, and in Chapter 7 to study the condition numbers for inversion of Fiedler matrices.

3.1 The inverse of a Fiedler Matrix

For $k = 1, \ldots, n-1$, the matrices M_k defined in (1.3) are nonsingular for any value of the coefficients a_k, while the matrix M_0 is nonsingular if and only if $a_0 \neq 0$. In this case, the inverses of these matrices are

$$M_0^{-1} = \begin{bmatrix} I_{n-1} & 0 \\ 0 & -1/a_0 \end{bmatrix}, \qquad M_k^{-1} = \begin{bmatrix} I_{n-k-1} & & & \\ & 0 & 1 & \\ & 1 & a_k & \\ & & & I_{k-1} \end{bmatrix}, \quad k = 1, 2, \ldots, n-1. \tag{3.1}$$

For any bijection σ, the Fiedler matrix M_σ in (2.1) is nonsingular if and only if $a_0 \neq 0$, that is, if $\lambda = 0$ is not a root of $p(z)$, and (3.1) allows us to obtain the following factorized expression of M_σ^{-1} given by

$$M_\sigma^{-1} = (M_{\sigma^{-1}(1)} \cdots M_{\sigma^{-1}(n)})^{-1} = M_{\sigma^{-1}(n)}^{-1} \ldots M_{\sigma^{-1}(1)}^{-1}.$$

However, as we did in `Algorithm 1` (Theorem 2.16) for M_σ, it is possible to construct the inverse of any Fiedler matrix using a similar algorithm, namely `Algorithm 2` in Theorem 3.1. This algorithm allows us to prove easily some key properties of M_σ^{-1} in Theorem 3.2. Note that `Algorithm 2` is not operation free, although the only arithmetic operations involved are multiplications of certain coefficients of $p(z)$ by $1/a_0$ (see Theorem 3.2).

Theorem 3.1. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. Then* `Algorithm 2` *constructs M_σ^{-1}.*

`Algorithm 2.` *Given* $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$, *with* $a_0 \neq 0$, *and a bijection* σ, *the following algorithm constructs* M_σ^{-1}.

if σ has a consecution at 0 then

$$B_0 = \begin{bmatrix} 0 & -1/a_0 \\ 1 & -a_1/a_0 \end{bmatrix}$$

else

$$B_0 = \begin{bmatrix} 0 & 1 \\ -1/a_0 & -a_1/a_0 \end{bmatrix}$$

endif

for $i = 1 : n-2$

if σ has a consecution at i then

$$B_i = \begin{bmatrix} 0 & B_{i-1}(1,:) \\ 1 & a_{i+1}B_{i-1}(1,:) \\ 0 & B_{i-1}(2:i+1,:) \end{bmatrix}$$

else

$$B_i = \begin{bmatrix} 0 & 1 & 0 \\ B_{i-1}(:,1) & a_{i+1}B_{i-1}(:,1) & B_{i-1}(:,2:i+1) \end{bmatrix}$$

endif

endfor

$M_\sigma^{-1} = B_{n-2}$.

Proof. Let $\{W_0, W_1, \ldots, W_{n-2}\}$ be the sequence of matrices constructed by `Algorithm 1`, in Theorem 2.16, and $\{B_0, B_1, \ldots, B_{n-2}\}$ be the sequence of matrices constructed by `Algorithm 2`. The proof consists of proving by induction that $W_i B_i = I_{i+2}$, i.e., that $B_i = W_i^{-1}$, which implies the theorem just by taking $i = n-2$.

If σ has a consecution at 0, then a direct multiplication of 2×2 matrices leads to $W_0 B_0 = I_2$. The same happens if σ has an inversion at 0. Let us assume that $W_{i-1}B_{i-1} = I_{i+1}$ for some $i - 1 \geq 0$ and let us prove $W_i B_i = I_{i+2}$. If σ has a consecution at i, then, from `Algorithms 1` and `2`, we get

$$W_i B_i = \begin{bmatrix} 1 & -a_{i+1}B_{i-1}(1,:) + a_{i+1}B_{i-1}(1,:) \\ 0 & W_{i-1}(:,1)B_{i-1}(1,:) + W_{i-1}(:,2:i+1)B_{i-1}(2:i+1,:) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & W_{i-1}B_{i-1} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & I_{i+1} \end{bmatrix}.$$

If σ has an inversion at i, then the proof is similar and is omitted. □

`Algorithm 2` allows us to easily get information on the entries of M_σ^{-1} in Theorem 3.2. The quantity t_σ, that is, the number of initial consecutions or inversions of a bijection σ (see Part (a) in Definition 2.8), will play a key role.

Theorem 3.2. *Let* $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ *with* $n \geq 2$ *and* $a_0 \neq 0$, *let* $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ *be a bijection, let* M_σ *be the Fiedler matrix of* $p(z)$ *associated with* σ, *and let* t_σ *be the number of initial consecutions or inversions of* σ. *Then:*

(a) M_σ^{-1} *has* $t_\sigma + 1$ *entries equal to* $-\dfrac{1}{a_0}, -\dfrac{a_1}{a_0}, \ldots, -\dfrac{a_{t_\sigma}}{a_0}$, *with exactly one copy of each.*

(b) M_σ^{-1} *has* $n - 1 - t_\sigma$ *entries equal to* $a_{t_\sigma+1}, a_{t_\sigma+2}, \ldots, a_{n-1}$, *with exactly one copy of each.*

(c) M_σ^{-1} *has* $n - 1$ *entries equal to* 1.

(d) *The rest of the entries of* M_σ^{-1} *are equal to* 0.

Proof. Recall that, according to Definition 2.8, $t_\sigma = \mathfrak{c}_0$ if $\mathfrak{c}_0 \neq 0$, and $t_\sigma = \mathfrak{i}_0$ if $\mathfrak{c}_0 = 0$. The case $\mathfrak{c}_0 = 0$ follows from the case $\mathfrak{c}_0 \neq 0$ by applying the result to the Fiedler matrix M_σ^T, which corresponds to a bijection with $\mathfrak{i}_0$ initial consecutions. Therefore, we prove only the result for $t_\sigma = \mathfrak{c}_0 \neq 0$. In this case, the bijection σ has consecutions at $0, 1, 2, \ldots, \mathfrak{c}_0 - 1$ and an inversion at $\mathfrak{c}_0$. Therefore, a direct application of `Algorithm 2` leads to

$$B_{\mathfrak{c}_0-1} = \begin{bmatrix} 0 & 0 & \ldots & & 0 & -1/a_0 \\ 1 & 0 & \ddots & & 0 & -a_{\mathfrak{c}_0}/a_0 \\ & 1 & \ddots & & & \vdots \\ & & \ddots & \ddots & & \vdots \\ & & & 1 & 0 & -a_2/a_0 \\ & & & & 1 & -a_1/a_0 \end{bmatrix}, \; B_{\mathfrak{c}_0} = \left[\begin{array}{c|cccccc} 0 & 1 & 0 & \ldots & & 0 & 0 \\ \hline 0 & 0 & 0 & \ldots & & 0 & -1/a_0 \\ 1 & a_{\mathfrak{c}_0+1} & 0 & \ddots & & 0 & -a_{\mathfrak{c}_0}/a_0 \\ 0 & 0 & 1 & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & \ddots & & \vdots \\ \vdots & \vdots & & & 1 & 0 & -a_2/a_0 \\ 0 & 0 & & & & 1 & -a_1/a_0 \end{array}\right]. \quad (3.2)$$

Observe that the nonzero entries of $B_{\mathfrak{c}_0}$ are: $\mathfrak{c}_0 + 1$ entries equal to 1, $-\frac{1}{a_0}, -\frac{a_1}{a_0}, \ldots, -\frac{a_{\mathfrak{c}_0}}{a_0}$, and $a_{\mathfrak{c}_0+1}$. In addition, both the first row and the first column of $B_{\mathfrak{c}_0}$ satisfy that they have only one nonzero entry and that this entry is equal to 1.

From `Algorithm 2`, one obtains by inspection the following property: if the first row and the first column of B_{i-1} satisfy that they have only one nonzero entry and that this entry is equal to 1, then (a) the nonzero entries of B_i are those of B_{i-1} together with an additional 1 and a_{i+1}, and (b) the first row and the first column of B_i have also only one nonzero entry and this entry is equal to 1.

This property and (3.2) imply that the nonzero entries of $M_\sigma^{-1} = B_{n-2}$ are those of $B_{\mathfrak{c}_0}$ together with $n - 2 - \mathfrak{c}_0$ entries equal to 1 and $a_{\mathfrak{c}_0+2}, a_{\mathfrak{c}_0+3}, \ldots, a_{n-1}$. This completes the proof. □

Next, we illustrate Theorem 3.2 with some examples of inverses of Fiedler matrices. In these examples it may be seen the dependence on t_σ of the nonzero entries of the inverse of a Fiedler matrix.

Example 3.3. Let $p(z) = z^7 + \sum_{k=0}^{6} a_k z^k$ be a monic polynomial of degree 7, and let C_2 and F be the second Frobenius companion matrix of $p(z)$ and the Fiedler matrix of $p(z)$ defined in (2.8), respectively. Then

$$C_2^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{a_0} \\ 1 & 0 & 0 & 0 & 0 & 0 & -\frac{a_6}{a_0} \\ 0 & 1 & 0 & 0 & 0 & 0 & -\frac{a_5}{a_0} \\ 0 & 0 & 1 & 0 & 0 & 0 & -\frac{a_4}{a_0} \\ 0 & 0 & 0 & 1 & 0 & 0 & -\frac{a_3}{a_0} \\ 0 & 0 & 0 & 0 & 1 & 0 & -\frac{a_2}{a_0} \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{a_1}{a_0} \end{bmatrix} \quad \text{and} \quad F^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & a_6 \\ 0 & 1 & 0 & 0 & 0 & 0 & a_5 \\ 0 & 0 & 1 & 0 & 0 & 0 & a_4 \\ 0 & 0 & 0 & 1 & 0 & 0 & a_3 \\ 0 & 0 & 0 & 0 & 1 & 0 & a_2 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{a_0} & -\frac{a_1}{a_0} \end{bmatrix}.$$

Recall that C_2 is a Fiedler matrix M_σ associated with a bijection σ such that $\mathrm{PCIS}(\sigma) = (1, 1, 1, 1, 1, 1)$ and $t_\sigma = 6$, and that F is a Fiedler matrix M_ρ associated with a bijection ρ such that $\mathrm{PCIS}(\rho) = (0, 1, 1, 1, 1, 1)$ and $t_\rho = 1$.

Example 3.4. Let $p(z) = z^7 + \sum_{k=0}^{6} a_k z^k$ be a monic polynomial of degree 7, and let P_1 and P_2

be the Fiedler matrices of $p(z)$ defined in (2.7). Then

$$P_1^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & a_6 & 0 & a_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & a_4 & 0 & a_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{a_0} \\ 0 & 0 & 0 & 0 & 1 & a_2 & -\frac{a_1}{a_0} \end{bmatrix} \quad \text{and} \quad P_2^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & a_6 & 0 & a_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & a_4 & 0 & a_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -\frac{1}{a_0} & -\frac{a_2}{a_0} & -\frac{a_1}{a_0} \end{bmatrix}.$$

Recall that P_1 is a Fiedler matrix M_σ associated with a bijection σ such that $\mathrm{PCIS}(\sigma) = (1,0,1,0,1,0)$ and $t_\sigma = 1$, and that P_2 is a Fiedler matrix M_ρ associated with a bijection ρ such that $\mathrm{PCIS}(\rho) = (0,0,1,0,1,0)$ and $t_\rho = 2$.

3.2 Formulas for the $\infty-$norm, $1-$norm and Frobenius norm of Fiedler matrices

Theorems 2.19 and 3.2 describe which are the non-identically zero entries of a Fiedler matrix M_σ and its inverse M_σ^{-1}. This information is enough for getting the Frobenius norms of Fiedler matrices and their inverses.

Corollary 3.5. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$, let $\sigma : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be a bijection, let M_σ be the Fiedler matrix of $p(z)$ associated with σ, and let t_σ be the number of initial consecutions or inversions of σ. Then:*

$$\|M_\sigma\|_F^2 = (n-1) + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2, \tag{3.3}$$

which is independent on σ and depends only on $p(z)$, and

$$\|M_\sigma^{-1}\|_F^2 = (n-1) + \frac{1 + |a_1|^2 + \cdots + |a_{t_\sigma}|^2}{|a_0|^2} + |a_{t_\sigma+1}|^2 + \cdots + |a_{n-1}|^2. \tag{3.4}$$

In contrast with $\|M_\sigma\|_F$, the quantity $\|M_\sigma^{-1}\|_F$ depends on σ, but only through the number of its initial consecutions or inversions t_σ. This implies that very different Fiedler matrices can have inverses with the same Frobenius norm, see, for example, F^{-1} and P_1^{-1} in Examples 3.3 and 3.4.

Theorems 2.19 and 3.2 do not give information on the positions where the non-identically zero entries of a Fiedler matrix and its inverse are placed in. In order to obtain expressions for $\|M_\sigma\|_\infty$ and $\|M_\sigma^{-1}\|_\infty$, we need to know how the non-identically zero entries of these two matrices are distributed by rows. This is presented in Lemma 3.6 for M_σ and in Lemma 3.7 for M_σ^{-1}. Once these two lemmas are known, we get easily Theorem 3.8, where the formulas for $\|M_\sigma\|_\infty$ and $\|M_\sigma^{-1}\|_\infty$ are finally stated. As a corollary of Theorem 3.8 and Theorem 2.12 the formulas for $\|M_\sigma\|_1$ and $\|M_\sigma^{-1}\|_1$ are obtained and presented in Theorem 3.9.

The results in this section require the *partial sums of the entries of* $\mathrm{CISS}(\sigma)$, that were previously used in [45, p. 2193]. We recall now their definitions: let $\sigma : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be a bijection and let $\mathrm{CISS}(\sigma) = (\mathfrak{c}_0, \mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{c}_\ell, \mathfrak{i}_\ell)$ be the consecution-inversion structure sequence of σ introduced in Definition 2.4. Then

$$s_k := \sum_{j=0}^{k} (\mathfrak{c}_j + \mathfrak{i}_j), \quad \text{for } k = 0,1,\ldots,\ell, \qquad s_{-1} := 0. \tag{3.5}$$

Observe that $s_\ell = n-1$ is the total number of consecutions and inversions of σ, that if $\mathfrak{c}_0 = 0$ then $s_0 = \mathfrak{i}_0$, and that $s_k = s_{k-1} + \mathfrak{c}_k + \mathfrak{i}_k$, for $k = 0,1,\ldots,\ell$.

Lemma 3.6. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let $\mathrm{CISS}(\sigma) = (\mathfrak{c}_0, \mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{c}_\ell, \mathfrak{i}_\ell)$ be the consecution-inversion structure sequence of σ, and let $\{s_k\}_{k=-1}^{\ell}$ be the partial sums of the entries of $\mathrm{CISS}(\sigma)$. If M_σ is the Fiedler matrix of $p(z)$ associated with σ, then the non-zero entries of M_σ are placed as specified in the following statements.*

(a) *Each of the $(n-1)$ entries equal to 1 is in a different row of M_σ. The only row of M_σ which does not contain an entry equal to 1 is*

  (i) *the nth row, if $\mathfrak{c}_0 > 0$;*

  (ii) *the $(n - \mathfrak{i}_0)$th row, if $\mathfrak{c}_0 = 0$.*

(b) *The entries $-a_0, -a_1, \ldots, -a_{s_0}$ of M_σ satisfy:*

  (i) *If $\mathfrak{c}_0 > 0$, then*

  - *each of the entries $-a_0, -a_1, \ldots, -a_{\mathfrak{c}_0 - 1}$ is in a different row of M_σ, each of these rows does not contain any other entry equal to $-a_i$, for $i = 0, 1, \ldots, n-1$, and $-a_0$ is in the nth row; and*
  - *the entries $-a_{\mathfrak{c}_0}, -a_{\mathfrak{c}_0+1}, \ldots, -a_{\mathfrak{c}_0+\mathfrak{i}_0}$ are all of them in the same row of M_σ and this row does not contain any other entry equal to $-a_i$, for $i = 0, 1, \ldots, n-1$.*

  (ii) *If $\mathfrak{c}_0 = 0$, then the entries $-a_0, -a_1, \ldots, -a_{\mathfrak{i}_0}$ are all of them in the $(n - \mathfrak{i}_0)$th row of M_σ and these are the only non-zero entries in this row.*

(c) *For each $k = 1, \ldots, \ell$, the entries $-a_{s_{k-1}+1}, -a_{s_{k-1}+2}, \ldots, -a_{s_k}$ of M_σ satisfy:*

  - *each of the entries $-a_{s_{k-1}+1}, \ldots, -a_{s_{k-1}+\mathfrak{c}_k-1}$ is in a different row of M_σ and each of these rows does not contain any other entry equal to $-a_i$, for $i = 0, 1, \ldots, n-1$; and*
  - *the entries $-a_{s_{k-1}+\mathfrak{c}_k}, -a_{s_{k-1}+\mathfrak{c}_k+1}, \ldots, -a_{s_k}$ are all of them in the same row of M_σ and this row does not contain any other entry equal to $-a_i$, for $i = 0, 1, \ldots, n-1$.*

Proof. The formal proof follows an inductive argument based on `Algorithm 1` in Theorem 2.16. We only sketch the main idea since the details to complete the proof are straightforward but somewhat long. The result is obviously true for the matrix W_0 appearing in `Algorithm 1`, or in other words, is obviously true for polynomials of degree 2. Then the induction hypothesis is that Lemma 3.6 holds for W_{n-3}, or in other words for polynomials of degree $n-1$, and then the way `Algorithm 1` constructs $W_{n-2} = M_\sigma$ from W_{n-2} is used to prove that the entries of M_σ satisfy Lemma 3.6. For this purpose, four cases should be considered, depending on whether σ has a consecution or an inversion at $n-3$, and on whether σ has a consecution or an inversion at $n-2$. □

Next, we determine in Lemma 3.7 the distribution by rows of the non-zero entries of M_σ^{-1}.

Lemma 3.7. *With the same notation and hypotheses that in Lemma 3.6, let us assume in addition that $a_0 \neq 0$ and that t_σ is the number of initial consecutions or inversions of σ. Then the non-zero entries of M_σ^{-1} are placed as specified in the following statements.*

(a) *Each of the $(n-1)$ entries equal to 1 is in a different row of M_σ^{-1}. The only row of M_σ^{-1} which does not contain an entry equal to 1 is the $(n - \mathfrak{c}_0)$th row.*

(b) *The entries $-\dfrac{1}{a_0}, -\dfrac{a_1}{a_0}, \ldots, -\dfrac{a_{t_\sigma}}{a_0}, a_{t_\sigma+1}, a_{t_\sigma+2}, \ldots, a_{s_0}$ of M_σ^{-1} satisfy*[1]

[1] Observe that, if $\mathfrak{c}_0 = 0$, then there are no entries $a_{t_\sigma+1}, a_{t_\sigma+2}, \ldots, a_{s_0}$ since $s_0 = \mathfrak{i}_0 = t_\sigma$.

(i) *If* $\mathfrak{c}_0 > 0$, *then*

- $-1/a_0$ *is the only non-zero entry in the* $(n-\mathfrak{c}_0)$*th row of* M_σ^{-1};
- *if, in addition,* $\mathfrak{c}_0 > 1$, *then each of the entries* $-a_1/a_0, \ldots, -a_{\mathfrak{c}_0-1}/a_0$ *is in a different row of* M_σ^{-1} *and each of these rows does not contain any other entry of the set* $\{-1/a_0, -a_1/a_0, \ldots, -a_{\mathfrak{c}_0}/a_0, a_{\mathfrak{c}_0+1}, \ldots, a_{n-1}\}$;
- *the entries* $-a_{\mathfrak{c}_0}/a_0, a_{\mathfrak{c}_0+1}, \ldots, a_{s_0}$ *are all of them in the same row of* M_σ^{-1} *and this row does not contain any other entry of the set* $\{-1/a_0, -a_1/a_0, \ldots, -a_{\mathfrak{c}_0}/a_0, a_{\mathfrak{c}_0+1}, \ldots, a_{n-1}\}$.

(ii) *If* $\mathfrak{c}_0 = 0$, *then the entries* $-1/a_0, -a_1/a_0, \ldots, -a_{\mathfrak{i}_0}/a_0$ *are all of them in the* n*th row of* M_σ^{-1} *and these are the only non-zero entries in this row.*

(c) *For each* $k = 1, \ldots, \ell$, *the entries* $a_{s_{k-1}+1}, a_{s_{k-1}+2}, \ldots, a_{s_k}$ *of* M_σ^{-1} *satisfy:*

- *each of the entries* $a_{s_{k-1}+1}, \ldots, a_{s_{k-1}+\mathfrak{c}_k-1}$ *is in a different row of* M_σ^{-1} *and each of these rows does not contain any other entry of the set* $\{-1/a_0, -a_1/a_0, \ldots, -a_{t_\sigma}/a_0, a_{t_\sigma+1}, \ldots, a_{n-1}\}$; *and*
- *the entries* $a_{s_{k-1}+\mathfrak{c}_k}, a_{s_{k-1}+\mathfrak{c}_k+1}, \ldots, a_{s_k}$ *are all of them in the same row of* M_σ^{-1} *and this row does not contain any other entry of the set* $\{-1/a_0, -a_1/a_0, \ldots, -a_{t_\sigma}/a_0, a_{t_\sigma+1}, \ldots, a_{n-1}\}$.

Proof. The proof is similar to the one of Lemma 3.6 but using the matrices B_i appearing in `Algorithm 1` instead of the matrices W_i. □

Lemmas 3.6 and 3.7 allow us to prove easily the main result in this section, that is, Theorem 3.8. The simple proof is omitted.

Theorem 3.8. *Let* $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ *be a monic polynomial with degree* $n \geq 2$, *let* $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ *be a bijection, let* $\mathrm{CISS}(\sigma) = (\mathfrak{c}_0, \mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{c}_\ell, \mathfrak{i}_\ell)$ *be the consecution-inversion structure sequence of* σ, *let* s_k, *for* $k = 0, 1, \ldots, \ell$, *be the partial sums defined in* (3.5), *and let* M_σ *be the Fiedler matrix of* $p(z)$ *associated with* σ. *Let us define the quantities*

$$\gamma_{\sigma,0}(p) = \begin{cases} \max\{1 + |a_1|, \ldots, 1 + |a_{\mathfrak{c}_0-1}|, 1 + |a_{\mathfrak{c}_0}| + |a_{\mathfrak{c}_0+1}| + \cdots + |a_{s_0}|\}, & \text{if } \mathfrak{c}_0 > 0, \\ \max\{|a_0| + |a_1| + \cdots + |a_{s_0}|\,, 1\}, & \text{if } \mathfrak{c}_0 = 0, \end{cases}$$

if $a_0 \neq 0$, *also the quantities*

$$\delta_\sigma(p) = \begin{cases} \max\left\{1 + \frac{|a_1|}{|a_0|}, \ldots, 1 + \frac{|a_{\mathfrak{c}_0-1}|}{|a_0|}, 1 + \frac{|a_{\mathfrak{c}_0}|}{|a_0|} + |a_{\mathfrak{c}_0+1}| + \cdots + |a_{s_0}|\right\}, & \text{if } \mathfrak{c}_0 > 0, \\ \max\left\{\frac{1}{|a_0|} + \frac{|a_1|}{|a_0|} + \cdots + \frac{|a_{s_0}|}{|a_0|}\,, 1\right\}, & \text{if } \mathfrak{c}_0 = 0, \end{cases}$$

and finally, for $k = 1, \ldots, \ell$, *the quantities*

$$\gamma_{\sigma,k}(p) = \max\{1 + |a_{s_{k-1}+1}|, \ldots, 1 + |a_{s_{k-1}+\mathfrak{c}_k-1}|, 1 + |a_{s_{k-1}+\mathfrak{c}_k}| + \cdots + |a_{s_k}|\},$$

where, if $\mathfrak{c}_k = 1$, *for some* $k = 0, 1, \ldots, \ell$, *then the first* $\mathfrak{c}_k - 1$ *terms within the maximums defining* $\gamma_{\sigma,k}(p)$ *or* $\delta_\sigma(p)$ *do not appear. Then*

$$\|M_\sigma\|_\infty = \max\{|a_0|, \gamma_{\sigma,0}(p), \gamma_{\sigma,1}(p), \ldots, \gamma_{\sigma,\ell}(p)\}, \tag{3.6}$$

and

$$\|M_\sigma^{-1}\|_\infty = \max\left\{\frac{1}{|a_0|}, \delta_\sigma(p), \gamma_{\sigma,1}(p), \ldots, \gamma_{\sigma,\ell}(p)\right\}. \tag{3.7}$$

As an immediate consequence of Theorems 2.12 and 3.8 we get formulas for $\|M_\sigma\|_1$ and $\|M_\sigma^{-1}\|_1$.

Theorem 3.9. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. Then,*

$$\|M_\sigma\|_1 = \|M_\sigma^T\|_\infty = \|M_{\mathrm{rev}(\sigma)}\|_\infty,$$

and, if $a_0 \neq 0$,

$$\|M_\sigma^{-1}\|_1 = \| \left(M_\sigma^{-1}\right)^T \|_\infty = \|M_{\mathrm{rev}(\sigma)}^{-1}\|_\infty,$$

where $\mathrm{rev}(\sigma)$ is the reversal bijection of σ.

Chapter 4

Singular values of Fiedler matrices

We have mentioned in Section 1.3 that in [95] (see also [99]), the authors prove that the Frobenius companion matrices associated with a monic polynomial $p(z)$ as in (1.1) have $n-2$ singular values equal to 1 and that the largest and the smallest singular values are the square roots of (1.29). The reason behind these properties is that C_1 (and C_2) can be written as a sum of a unitary matrix plus a rank-one matrix as follows

$$C_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_1 & -a_0 - 1 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \end{bmatrix}. \tag{4.1}$$

This expression immediately allows us to prove that C_1 has at least $n-2$ singular values equal to 1 and that the squares of the remaining two singular values can be obtained as the eigenvalues of a simple 2×2 matrix (we will present in Lemma 4.25 a general version of this result). In fact, the unitary matrix in the sum (4.1) has an additional property: it is a *permutation matrix*, i.e., a matrix obtained by permuting the rows (or columns) of the identity matrix.

Fiedler matrices different from the Frobenius companion matrices cannot be expressed as "unitary plus rank-one matrices", but we will see in Section 4.2 that every Fiedler matrix of $p(z)$ can be expressed as a sum of a permutation matrix plus a matrix whose rank varies from 1 to $\lfloor (n+1)/2 \rfloor$. In addition, we will show how to construct these two summands via simple algorithms and how to determine the rank of the second summand. In plain words, this will imply that many Fiedler matrices admit expressions as "unitary plus low-rank matrices" and so have a certain number of singular values equal to 1. We illustrate in Example 4.1 these ideas.

Example 4.1. We consider monic polynomials with degree 8, i.e., $p(z) = z^8 + \sum_{k=0}^{7} a_k z^k$.

1. We consider first the pentadiagonal Fiedler matrix P_1 in (2.7). It can be written as

$$P_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} -a_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -a_6 & 0 & -a_5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_4 & 0 & -a_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -a_2 & 0 & -a_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -a_0 - 1 & 0 \end{bmatrix}. \tag{4.2}$$

The first summand is a permutation matrix and the second one has rank at most $4 = \lfloor (n+1)/2 \rfloor$ (to see this, perform Gaussian elimination by rows). In fact, if $a_i \neq 0$ for $i = 1, \ldots, 7$, then the rank is exactly 4.

2. The second example corresponds to a Fiedler matrix with $\mathrm{CISS}(\sigma) = (4, 3)$. It can be expressed as

$$M_\sigma = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} -a_7 & -a_6 & -a_5 & -a_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_0 - 1 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{4.3}$$

The first summand is again a permutation matrix and the second one has rank at most 2. In fact, if $a_i \neq 0$ for $i = 1, \ldots, 7$, then the rank is exactly 2. `Algorithm 1` in Theorem 2.16 allows the reader to easily check that these properties hold for Fiedler matrices of polynomials with arbitrary degree $n \geq 2$ associated with bijections that have all their consecutions in consecutive indices and all their inversions in consecutive indices, that is, those with $\mathrm{CISS}(\sigma) = (\mathfrak{c}_0, \mathfrak{i}_0)$, $\mathfrak{c}_0 \neq 0$ and $\mathfrak{i}_0 \neq 0$, or those with $\mathrm{CISS}(\sigma) = (0, \mathfrak{i}_0, \mathfrak{c}_1, 0)$, $\mathfrak{i}_0 \neq 0$ and $\mathfrak{c}_1 \neq 0$.

Observe that, if all zero rows and columns are removed in the second summands in (4.2) and (4.3), then, in both cases, we get matrices whose nonzero entries follow a very special pattern:

$$\begin{bmatrix} -a_7 & 0 & 0 & 0 \\ -a_6 & -a_5 & 0 & 0 \\ 0 & -a_4 & -a_3 & 0 \\ 0 & 0 & -a_2 & -a_1 \\ 0 & 0 & 0 & -a_0 - 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -a_7 & -a_6 & -a_5 & -a_4 \\ 0 & 0 & 0 & -a_3 \\ 0 & 0 & 0 & -a_2 \\ 0 & 0 & 0 & -a_1 \\ 0 & 0 & 0 & -a_0 - 1 \end{bmatrix}.$$

These matrices are particular cases of a family of matrices defined in Section 4.1 and named *staircase matrices*. These matrices and the determination of their ranks will be the key point in our analysis of the singular values of Fiedler matrices. The study of staircase matrices is the subject of section 4.1.

Then, based on the study of staircase matrices, in Section 4.2 we study some properties of the singular values of Fiedler matrices. More precisely, we determine how many of their singular values are equal to one and, for those that are not, we show that they can be obtained from the square roots of the eigenvalues of certain matrices that may have a size much smaller than n and that are easily constructible from the coefficients of $p(z)$.

4.1 Staircase matrices

Staircase matrices are matrices whose nonzero entries follow a very special pattern. We assume throughout this section that these matrices have more than one row or more than one column to avoid the trivial 1×1 case that may complicate the definition.

Definition 4.2. *Given a matrix $A = (a_{ij}) \in \mathbb{C}^{m \times p}$, we say that A is a* staircase matrix *if A satisfies the following properties:*

1. *If $a_{i,j_1} \neq 0$ and $a_{i,j_2} \neq 0$, for some $1 \leq i \leq m$ and $1 \leq j_1 \leq j_2 \leq p$, then $a_{ij} \neq 0$ for all $j_1 \leq j \leq j_2$.*

2. *If* $a_{i1} = a_{i2} = \cdots = a_{i,j-1} = 0$ *and* $a_{ij} \neq 0$*, for some* $1 < i \leq m,\ 1 \leq j \leq p$*, then* $a_{i-1,j} \neq 0$ *and* $a_{i-1,j+1} = 0$*, whenever* $j + 1 \leq p$*.*

3. $a_{11} \neq 0$ *and* $a_{mp} \neq 0$*.*

A matrix $B = (b_{ij}) \in \mathbb{C}^{m\times p}$ *is said to be a* generalized staircase matrix *if it is obtained from a staircase matrix by turning some nonzero entries into zero entries.*

The first condition in Definition 4.2 means that all nonzero entries in a given row of A are placed in consecutive columns. The second condition means that the first nonzero entry in a given row of A is placed in the same column as the last nonzero entry of the immediate upper row.

Example 4.3. The following matrices are staircase matrices:

$$A = \begin{bmatrix} \times & \times & \times & & & & & \\ & & \times & \times & & & & \\ & & & \times & \times & \times & & \\ & & & & & \times & & \\ & & & & & \times & & \\ & & & & & \times & \times & \times \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} \times & \times & \times & & & & \\ & & \times & \times & & & \\ & & & \times & \times & \times & \\ & & & & & \times & \\ & & & & & \times & \\ & & & & & \times & \end{bmatrix},$$

where the symbol $\times$ denotes the nonzero entries (here and in all the examples of this section). A generalized staircase matrix compatible with A is obtained by replacing some of the $\times$ entries of A by 0. In plain words, one can say that generalized staircase matrices may have "holes" in the steps.

Notice that, as a consequence of the second and third conditions in Definition 4.2, every row and every column in a staircase matrix has at least one nonzero entry.

Definition 4.4. *Let* $A \in \mathbb{C}^{m\times p}$ *be a staircase matrix. We say that a nonzero entry* a_{ij} *is a* corner entry *of* A *if one (or both) of the following conditions holds:*

1. $i = j = 1$ *or* $i = m$ *and* $j = p$*.*

2. a_{ij} *is the first or the last nonzero entry in the* ith *row and there are more than one nonzero entries in the* ith *row.*

Example 4.5. *Let* A *and* C *be the staircase matrices in Example* 4.3*. Then*

$$A = \begin{bmatrix} \otimes & \times & \otimes & & & & & \\ & & \otimes & \otimes & & & & \\ & & & \otimes & \times & \otimes & & \\ & & & & & \times & & \\ & & & & & \times & & \\ & & & & & \otimes & \times & \otimes \end{bmatrix} \quad \textit{and} \quad C = \begin{bmatrix} \otimes & \times & \otimes & & & & \\ & & \otimes & \otimes & & & \\ & & & \otimes & \times & \otimes & \\ & & & & & \times & \\ & & & & & \times & \\ & & & & & \otimes & \end{bmatrix},$$

and the entries with the symbols $\otimes$ *are the corner entries.*

Definition 4.6. *Let* $A \in \mathbb{C}^{m\times p}$ *be a staircase matrix. We define the* ordered list of corner entries *of* A *as the ordered list* $(a_{i_1,j_1}, a_{i_2,j_2}, \ldots, a_{i_{t+1},j_{t+1}})$ *of all corner entries of* A*, where the corner entry* a_{i_r,j_r} *precedes the corner entry* a_{i_s,j_s} *if* $i_r < i_s$ *or* $i_r = i_s$ *and* $j_r < j_s$*.*

Example 4.7. *For the staircase matrices in Example* 4.5*, the ordered lists of corner entries of* A *and* C *are, respectively,* $\{a_{11}, a_{13}, a_{23}, a_{24}, a_{34}, a_{36}, a_{66}, a_{68}\}$ *and* $\{c_{11}, c_{13}, c_{23}, c_{24}, c_{34}, c_{36}, c_{66}\}$*.*

Notice that for two consecutive entries in the ordered list of corner entries of A, say a_{i_k,j_k} and $a_{i_{k+1},j_{k+1}}$, we always have $i_k = i_{k+1}$ or $j_k = j_{k+1}$ (but not both). This motivates the following definition.

Definition 4.8. *Let $A \in \mathbb{C}^{m\times p}$ be a staircase matrix and $(a_{i_1,j_1},\ldots,a_{i_{t+1},j_{t+1}})$ be the ordered list of corner entries of A. Then, for $1 \le k \le t$, the* kth flight *of A is the set of entries*

- $a_{i_k,j_k}, a_{i_k,j_k+1},\ldots,a_{i_k,j_{k+1}}$ *if* $i_k = i_{k+1}$, *or*
- $a_{i_k,j_k}, a_{i_k+1,j_k},\ldots,a_{i_{k+1},j_k}$ *if* $j_k = j_{k+1}$.

Notice that the number of flights of a staircase matrix A is equal to the number of corner entries of A minus one. We are particularly interested in the lengths of the flights of A. This notion is made precise in Definition 4.9.

Definition 4.9. *Let $A \in \mathbb{C}^{m\times p}$ be a staircase matrix and $(a_{i_1,j_1},\ldots,a_{i_{t+1},j_{t+1}})$ be the ordered list of corner entries of A. The* flight-length sequence *of A is the sequence*

$$\mathcal{F}(A) := (f_1, f_2, \ldots, f_t),$$

where $f_k = \max\{i_{k+1} - i_k, j_{k+1} - j_k\}$, for $k = 1,\ldots,t$.

We note that the kth term f_k in the flight-length sequence of A is equal to the number of entries in the kth flight of A minus one.

Definition 4.10. *Let $\mathbf{s} = (s_1, s_2, \ldots, s_t)$ be an ordered list of nonnegative integers. For each $j = 1, 2, \ldots, t$, we define the* length of the string of ones at the jth position *of* $\mathbf{s}$, *denoted by l_j, as*

- *a positive integer $l_j > 0$, if the following three conditions are satisfied:*
  - (i) $s_j = s_{j+1} = \cdots = s_{j+l_j-1} = 1$,
  - (ii) $s_{j-1} \neq 1$ *or* $j = 1$, *and*
  - (iii) $s_{j+l_j} \neq 1$ *or* $j + l_j - 1 = t$.
- $l_j = 0$, *otherwise.*

Let $(l_1, l_2, \ldots, l_t)$ be the ordered list of the lengths of the strings of ones of $\mathbf{s}$. *Then, the* list of positive lengths of the strings of ones of $\mathbf{s}$, *denoted by $\mathcal{L}(\mathbf{s})$, is the ordered list obtained from $(l_1, l_2, \ldots, l_t)$ after removing all zero entries. If* $\mathbf{s}$ *is a list containing no ones, then we set $\mathcal{L}(\mathbf{s}) := (0)$.*

Example 4.11. *For the list* $\mathbf{s} = (2, 1, 1, 1, 3, 1, 1, 2)$, *the list of the lengths of the strings of ones is* $(0, 3, 0, 0, 0, 2, 0, 0)$, *so we have* $\mathcal{L}(\mathbf{s}) = (3, 2)$.

Until now, we have not established any relationship between staircase matrices and Fiedler matrices. However, both types of matrices are closely connected in a way that will be shown in Section 4.2. In order to introduce this connection, we show in Theorem 4.12 that every staircase matrix with n nonzero entries can be constructed from the consecutions and inversions of a bijection $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$. Moreover, we show that the reduced consecution-inversion structure sequence of σ, RCISS(σ) (see part (b) in Definition 2.8), is the flight-length sequence of the matrix *in reversed order*. The reader is invited to focus on the similarities between `Algorithm 3` in Theorem 4.12 and `Algorithm 1` in Theorem 2.16, which will be exploited in depth in Section 4.2. However, note that in `Algorithm 3` we use the MATLAB notation $V(:, j : \texttt{end})$ to indicate the submatrix of V consisting of columns j through the last column (a similar notation is used for rows), because the sizes of the constructed matrices are not fixed. They depend on the number of consecutions and inversions of σ. In addition, if expressions like $V(:, 2 : 1)$ appear in `Algorithm 3`, then they should be understood as empty matrices. We warn also the reader that in `Algorithm 3` the staircase matrix is constructed starting from the lower-right entry, which may seem unnatural, but it is convenient for establishing the connection with Fiedler matrices and `Algorithm 1`.

Theorem 4.12. *Let $x_0, x_1, \ldots, x_{n-1}$ be $n \geq 2$ complex nonzero numbers, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and consider the following algorithm:*

> `Algorithm 3.` *Given $x_0, x_1, \ldots, x_{n-1}$ nonzero numbers and a bijection σ, the following algorithm constructs a matrix $\widetilde{V}_\sigma$ whose nonzero entries are precisely $x_0, x_1, \ldots, x_{n-1}$.*
>
> if σ has a consecution at 0 then
> $\quad \widetilde{V}_0 = \begin{bmatrix} x_1 \\ x_0 \end{bmatrix}$
> else
> $\quad \widetilde{V}_0 = \begin{bmatrix} x_1 & x_0 \end{bmatrix}$
> endif
> for $i = 1 : n-2$
> $\quad$ if σ has a consecution at i then
> $\quad\quad \widetilde{V}_i = \begin{bmatrix} x_{i+1} & 0 \\ \widetilde{V}_{i-1}(:,1) & \widetilde{V}_{i-1}(:, 2 : \texttt{end}) \end{bmatrix}$
> $\quad$ else
> $\quad\quad \widetilde{V}_i = \begin{bmatrix} x_{i+1} & \widetilde{V}_{i-1}(1,:) \\ 0 & \widetilde{V}_{i-1}(2 : \texttt{end}, :) \end{bmatrix}$
> $\quad$ endif
> endfor
> $\widetilde{V}_\sigma = \widetilde{V}_{n-2}$.

Then the matrix $\widetilde{V}_\sigma$ is a staircase matrix. Moreover, if $\mathrm{RCISS}(\sigma) = (p_1, p_2, \ldots, p_t)$ *is the reduced consecution-inversion structure sequence of σ, then the flight-length sequence of $\widetilde{V}_\sigma$ is $\mathcal{F}(\widetilde{V}_\sigma) = (p_t, p_{t-1}, \ldots, p_2, p_1)$.*

Conversely, given a staircase matrix A with n nonzero entries and flight-length sequence $\mathcal{F}(A) = (f_1, f_2, \ldots, f_t)$, there exists a bijection $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ such that $\mathrm{RCISS}(\sigma) = (f_t, \ldots, f_2, f_1)$ *and $A = \widetilde{V}_\sigma$, where $\widetilde{V}_\sigma$ is the output of* `Algorithm 3` *with the inputs σ and the list of the nonzero entries of A ordered from the lower-right to the upper-left entry*[1].

Proof. The proof is easy, so we only sketch the main points. In the proof we use a family of bijections $\sigma_i : \{0, 1, \ldots, i+1\} \to \{1, \ldots, i+2\}$, for $i = 0, 1, \ldots, n-2$, such that σ_i has a consecution (resp. inversion) at j, $0 \leq j \leq i$, if and only if σ has a consecution (resp. inversion) at j. Observe that $\widetilde{V}_i$ is constructed by applying `Algorithm 3` to the numbers $x_0, x_1, \ldots, x_{i+1}$ and the bijection σ_i. The bijection σ_{n-2} may be taken to be equal to σ.

Let us prove first the properties of $\widetilde{V}_\sigma$. It is obvious that $\widetilde{V}_0$ is a staircase matrix that has $\mathcal{F}(\widetilde{V}_0) = (1)$, and that $\mathrm{RCISS}(\sigma_0) = (1)$. Next, we proceed by induction. Assume that $\widetilde{V}_{i-1}$, for some $i - 1 \geq 0$, is a staircase matrix that has $\mathcal{F}(\widetilde{V}_{i-1}) = (p'_s, p_{s-1}, \ldots, p_2, p_1)$, where $(p_1, p_2, \ldots, p_{s-1}, p'_s) = \mathrm{RCISS}(\sigma_{i-1})$. Then the structure of `Algorithm 3` makes obvious that $\widetilde{V}_i$ is also a staircase matrix. Also if σ has two consecutions or two inversions at $i-1$ and i, then $\mathcal{F}(\widetilde{V}_i) = (p'_s + 1, p_{s-1}, \ldots, p_2, p_1)$ and also $(p_1, p_2, \ldots, p_{s-1}, p'_s + 1) = \mathrm{RCISS}(\sigma_i)$. If σ has a consecution at $i-1$ and an inversion at i, or viceversa, then $\mathcal{F}(\widetilde{V}_i) = (1, p'_s, p_{s-1}, \ldots, p_2, p_1)$ and also $(p_1, p_2, \ldots, p_{s-1}, p'_s, 1) = \mathrm{RCISS}(\sigma_i)$ (note that in this case $p'_s = p_s$). Therefore, the result is true for $\widetilde{V}_i$. The result in the statement follows by taking $i = n - 2$.

The "converse statement" is also immediate just by looking carefully at `Algorithm 3` and the reader is invited to complete the details. The only point to be remarked is that σ is not determined only by $\mathrm{RCISS}(\sigma)$. It is needed to also know whether σ has a consecution or an inversion at 0. Note that if the last flight of A, with length f_t, is an horizontal flight, i.e., it corresponds to entries

[1] More precisely, this order corresponds to list all the flights of A from 1 to t, to remove repeated entries, and to reverse the order of the obtained list.

in the same row, then σ has inversions at $0, 1, \ldots, f_t - 1$. On the contrary, if the last flight of A is a vertical flight, i.e., it corresponds to entries in the same column, then σ has consecutions at $0, 1, \ldots, f_t - 1$. □

Given n ordered nonzero numbers $x_0, x_1, \ldots, x_{n-1}$, Theorem 4.12 establishes a correspondence between bijections $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ and staircase matrices A that have as nonzero entries $x_0, x_1, \ldots, x_{n-1}$, with x_0 being the lower-right entry and x_{n-1} being the upper-left. Taking into account the relationship between $\mathrm{RCISS}(\sigma)$ and $\mathcal{F}(A)$ in Theorem 4.12, Definition 4.13 introduces, from two different but equivalent perspectives, the list that will allow us to determine the rank of a staircase matrix in Theorem 4.16.

Definition 4.13. (a) *Let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, 2, \ldots, n\}$ with $n \geq 2$ be a bijection and let $\mathrm{RCISS}(\sigma) = (p_1, p_2, \ldots, p_t)$ be the reduced consecution-inversion structure sequence of σ. Let $\mathbf{s} := (p_{t-1}, \ldots, p_2)$. Then, the* rank-determining list of σ, *denoted by $\mathcal{L}(\sigma)$, is $\mathcal{L}(\mathbf{s})$, that is, the list of positive lengths of the strings of ones of $\mathbf{s}$ introduced in Definition* 4.10. *If $\mathrm{RCISS}(\sigma)$ has* 1 *or* 2 *entries, then $\mathbf{s}$ is empty and we set $\mathcal{L}(\sigma) := (0)$.*

(b) *Let A be a staircase matrix and let $\mathcal{F}(A) = (f_1, f_2, \ldots, f_t)$ be its flight-length sequence. Let $\mathbf{s} := (f_2, \ldots, f_{t-1})$. Then, the* rank-determining list of A, *denoted by $\mathcal{L}(A)$, is $\mathcal{L}(\mathbf{s})$, that is, the list of positive lengths of the strings of ones of $\mathbf{s}$ introduced in Definition* 4.10. *If $\mathcal{F}(A)$ has* 1 *or* 2 *entries, then $\mathbf{s}$ is empty and we set $\mathcal{L}(A) := (0)$.*

Observe that $\mathcal{L}(\sigma)$ has been defined without any reference to staircase matrices. It depends only on the bijection σ. However, if $\widetilde{V}_\sigma$ is *any* staircase matrix constructed by `Algorithm 3` for this bijection, then $\mathcal{L}(\sigma) = \mathcal{L}(\widetilde{V}_\sigma)$, as a consequence of Theorem 4.12. Also, given any staircase matrix A, Theorem 4.12 guarantees that there exists a bijection σ such that $\mathcal{L}(A) = \mathcal{L}(\sigma)$. Therefore we will use the notation $\mathcal{L}(\sigma)$ or $\mathcal{L}(A)$ depending on which is more convenient for the specific result we are considering. We illustrate these concepts in Example 4.14.

Example 4.14. *The staircase matrix A in Example* 4.3 *has $\mathcal{F}(A) = (2, 1, 1, 1, 2, 3, 2)$ as flight-length sequence. Then A can be constructed by* `Algorithm 3` *via a bijection $\sigma : \{0, 1, \ldots, 12\} \to \{1, \ldots, 13\}$ with $\mathrm{RCISS}(\sigma) = (2, 3, 2, 1, 1, 1, 2)$. Note that the first two entries of $\mathrm{CISS}(\sigma)$ are $\mathfrak{c}_0 = 0$ and $\mathfrak{i}_0 = 2$, since the last flight of A corresponds to entries in the same row. Moreover, we have $\mathcal{L}(A) = \mathcal{L}(\sigma) = (3)$.*

Theorem 4.16 is the key result of this section. It gives the simple formula (4.4) for the rank of any staircase matrix in terms of its flight-length sequence. The formula (4.4) shows that to determine the rank of a staircase matrix is not completely trivial. The idea of the proof is to perform *Gaussian Elimination by rows and columns* starting from the upper-left corner. We illustrate this procedure in a simple case in Example 4.15, and then we state Theorem 4.16.

Example 4.15. *By Gaussian elimination, it is easy to determine the rank of any staircase matrix. Consider the matrix A in Example 4.3. Using elementary row and column replacement operations starting from the upper-left entry, we can transform the matrix A into*

$$A = \begin{bmatrix} \times & \times & \times & & & & & \\ & & \times & \times & & & & \\ & & & \times & \times & \times & & \\ & & & & & \times & & \\ & & & & & \times & & \\ & & & & & \times & \times & \times \end{bmatrix} \sim \begin{bmatrix} \times & 0 & 0 & & & & & \\ & & \times & 0 & & & & \\ & & & \times & 0 & 0 & & \\ & & & & & \times & & \\ & & & & & 0 & & \\ & & & & & 0 & \times & 0 \end{bmatrix}.$$

Hence $\operatorname{rank} A = 5$. *As can be seen in Theorem* 4.16, *the rank of a staircase matrix A can be obtained from the number of flights of A and the sequence $\mathcal{L}(A)$. It is important to notice the role played by those flights of length* 1 *different from the first and the last flights.*

In the rest of this chapter, given a real number x, we will use the standard notation $\lceil x \rceil$ for the smallest integer which is greater than or equal to x.

Theorem 4.16. *Let $A \in \mathbb{C}^{m\times p}$ be a staircase matrix, let $\mathcal{F}(A) = (f_1, f_2, \ldots, f_t)$ be the flight-length sequence of A, and let $\mathcal{L}(A) = (l_1, l_2, \ldots, l_q)$ be the rank-determining list of A. Then the rank of A is equal to*

$$\operatorname{rank} A = t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil . \tag{4.4}$$

Proof. The proof proceeds by induction on the number of flights t. For $t = 1$, the result is obviously true because all staircase matrices with only one flight have $\operatorname{rank} A = 1$ and $\mathcal{L}(A) = (0)$, so

$$t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil = 1 = \operatorname{rank} A.$$

By a similar argument the result is also true for $t = 2$, since in this case $\operatorname{rank} A = 2$ and $\mathcal{L}(A) = (0)$. Now, let us assume that the result is true for any staircase matrix with $t - 1 \geq 2$ flights. Let A and $\widehat{A}$ be staircase matrices with $\mathcal{F}(A) = (f_1, f_2, \ldots, f_t)$ and $\mathcal{F}(\widehat{A}) = (f_1, f_2, \ldots, f_{t-1})$ and such that A is obtained from $\widehat{A}$ by adding one flight with length f_t. Note that A and $\widehat{A}$ have different sizes. We distinguish two cases.

Case 1: $f_{t-1} \neq 1$. In fact, according to Definition 4.9 this means $f_{t-1} > 1$. In this case, $\mathcal{L}(A) = \mathcal{L}(\widehat{A}) = (l_1, l_2, \ldots, l_q)$. The reason is that $\mathcal{L}(A)$ is determined by the strings of ones in $(f_2, \ldots, f_{t-1})$, while $\mathcal{L}(\widehat{A})$ is determined by the strings of ones in $(f_2, \ldots, f_{t-2})$, and in both cases the strings of ones are the same.

In addition, $\operatorname{rank} A = 1 + \operatorname{rank} \widehat{A}$. To see this, assume without loss of generality that the last flight of A has all its entries in the same row (otherwise we transpose the matrix, which preserves the rank and the flight-length sequence). Therefore, A has more columns than $\widehat{A}$ and the same number of rows, i.e., $A \in \mathbb{C}^{m\times p}$ and $\widehat{A} \in \mathbb{C}^{m\times \ell}$ with $\ell < p$, and the last flight of $\widehat{A}$ has all its entries in the same column. This and the fact $f_{t-1} > 1$ mean that the last two rows of A are

$$\begin{aligned} A(m-1:m,:) &= \left[\begin{array}{cccc|ccc} 0 & \cdots & 0 & \times & 0 & \ldots & 0 \\ 0 & \cdots & 0 & \times & \times & \ldots & \times \end{array}\right] \sim \left[\begin{array}{cccc|ccc} 0 & \cdots & 0 & \times & 0 & \ldots & 0 \\ 0 & \cdots & 0 & 0 & \times & \ldots & \times \end{array}\right] \\ &= A'(m-1:m,:), \end{aligned} \tag{4.5}$$

where the symbol $\times$ denotes nonzero entries, the vertical line separates $\widehat{A}$ from those columns of A that are not columns of $\widehat{A}$, and we have performed an elementary row replacement operation to get A'. Since $A(1 : m-1, \ell+1 : p) = 0$, (4.5) implies that $\operatorname{rank} A = \operatorname{rank} A' = 1 + \operatorname{rank} \widehat{A}' = 1 + \operatorname{rank} \widehat{A}$.

The above equalities $\mathcal{L}(A) = \mathcal{L}(\widehat{A})$ and $\operatorname{rank} A = 1 + \operatorname{rank} \widehat{A}$, and the induction hypothesis imply

$$\operatorname{rank} A = \operatorname{rank} \widehat{A} + 1 = t - 1 - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil + 1 = t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil ,$$

which proves the result for A in **Case 1**.

Case 2: $f_{t-1} = 1$. In this case $\mathcal{L}(A) = (l_1, l_2, \ldots, l_q)$ with $l_q \geq 1$, and we need to distinguish two subcases: l_q even and l_q odd.

Case 2.1: $l_q = 2k$, with $k > 0$ an integer. In this case $l_q > 1$ and $\mathcal{L}(\widehat{A}) = (l_1, l_2, \ldots, l_{q-1}, l_q - 1)$. Definition 4.13 and $l_q = 2k$ imply

$$f_{t-2k} = f_{t-(2k-1)} = \cdots = f_{t-1} = 1, \quad \text{with } t - 2k \geq 2, \text{ and } f_{t-2k-1} > 1 \text{ if } t - 2k > 2. \tag{4.6}$$

Assume, as in **Case 1**, that the last flight of A has all its entries in the same row, which implies that $A \in \mathbb{C}^{m\times p}$ and $\widehat{A} \in \mathbb{C}^{m\times \ell}$ with $\ell < p$, and also that the last flight of $\widehat{A}$ has its two entries in the same column. This and (4.6) imply that if $t - 2k > 2$

$$A(m-k-1:m,:) = \left[\begin{array}{ccccccccc|ccc} 0 & \cdots & 0 & \times & & & & & & 0 & \dots & 0 \\ \vdots & & \vdots & \times & \times & & & & & \vdots & & \vdots \\ \vdots & & \vdots & & \times & \times & & & & \vdots & & \vdots \\ \vdots & & \vdots & & & \ddots & \ddots & & & \vdots & & \vdots \\ \vdots & & \vdots & & & & \times & \times & & 0 & \dots & 0 \\ 0 & \cdots & 0 & & & & & \times & & \times & \dots & \times \end{array}\right], \tag{4.7}$$

where the vertical line separates $\widehat{A}$ from those columns of A that are not columns of $\widehat{A}$. If we perform elementary row replacement operations in $A(m-k-1:m,:)$ starting from the top we get

$$A(m-k-1:m,:) \sim \left[\begin{array}{ccccccccc|ccc} 0 & \cdots & 0 & \times & & & & & & 0 & \dots & 0 \\ \vdots & & \vdots & 0 & \times & & & & & \vdots & & \vdots \\ \vdots & & \vdots & & 0 & \times & & & & \vdots & & \vdots \\ \vdots & & \vdots & & & \ddots & \ddots & & & \vdots & & \vdots \\ \vdots & & \vdots & & & & 0 & \times & & 0 & \dots & 0 \\ 0 & \cdots & 0 & & & & & 0 & & \times & \dots & \times \end{array}\right], \tag{4.8}$$

which implies

$$\operatorname{rank} A = 1 + \operatorname{rank} \widehat{A}. \tag{4.9}$$

If $t-2k = 2$, then $A(m-k-1:m,:)$ is as in (4.7) but removing all the left-most columns of zeros. So we also get (4.9).

The equalities $\mathcal{L}(\widehat{A}) = (l_1, l_2, \dots, l_{q-1}, l_q - 1)$, (4.9), and the induction hypothesis imply

$$\operatorname{rank} A = \operatorname{rank} \widehat{A} + 1 = t - 1 - \sum_{j=1}^{q-1} \left\lceil \frac{l_j}{2} \right\rceil - \left\lceil \frac{2k-1}{2} \right\rceil + 1 = t - \sum_{j=1}^{q-1} \left\lceil \frac{l_j}{2} \right\rceil - \left\lceil \frac{2k}{2} \right\rceil = t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil,$$

which proves the result for A in *Case 2.1.*

Case 2.2: $l_q = 2k+1$, with $k \geq 0$ an integer. The proof is similar to the one of *Case 2.1*, so we only emphasize the main differences and omit the details. To begin with, in this case

$$\mathcal{L}(\widehat{A}) = \begin{cases} (l_1, l_2, \dots, l_{q-1}, l_q - 1), & \text{if } l_q > 1, \\ (l_1, l_2, \dots, l_{q-1}), & \text{if } l_q = 1. \end{cases},$$

and one has to distinguish the cases $l_q > 1$ and $l_q = 1$. In both of them, it is satisfied that

$$\operatorname{rank} A = \operatorname{rank} \widehat{A}. \tag{4.10}$$

This follows because in this case the structure of A is

$$A(m-k-2:m,:) = \left[\begin{array}{ccccccc|ccc} \cdots & \cdots & 0 & & & & & 0 & \dots & 0 \\ \cdots & \cdots & \times & \times & & & & 0 & \dots & 0 \\ 0 & & 0 & \times & \times & & & \vdots & & \vdots \\ \vdots & & \vdots & & \times & \times & & \vdots & & \vdots \\ \vdots & & \vdots & & & \ddots & \ddots & \vdots & & \vdots \\ \vdots & & \vdots & & & & \times \quad \times & 0 & \dots & 0 \\ 0 & \cdots & 0 & & & & \times & \times & \dots & \times \end{array}\right], \tag{4.11}$$

and elementary column replacement operations starting from the left-most $\times$ entry shown in (4.11) allows us to make zeros all the entries to the right of the vertical line.

The equalities $\mathcal{L}(\widehat{A}) = (l_1, l_2, \dots, l_{q-1}, l_q - 1)$ (if $l_q > 1$), (4.10), and the induction hypothesis imply

$$\begin{aligned} \operatorname{rank} A = \operatorname{rank} \widehat{A} &= t - 1 - \sum_{j=1}^{q-1} \left\lceil \frac{l_j}{2} \right\rceil - \left\lceil \frac{2k}{2} \right\rceil = t - \sum_{j=1}^{q-1} \left\lceil \frac{l_j}{2} \right\rceil - (k+1) = \\ &= t - \sum_{j=1}^{q-1} \left\lceil \frac{l_j}{2} \right\rceil - \left\lceil \frac{2k+1}{2} \right\rceil = t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil, \end{aligned}$$

which proves the result for A in *Case 2.2*. Observe that the case $l_q = 1$ follows by taking $k = 0$ in the equation above. □

Theorem 4.16 shows, in particular, that the rank of a staircase matrix is not an increasing function of the number of flights, as it might be thought at a first glance, since intermediate flights of length 1 also affects the rank. Example 4.17 illustrates this fact.

Example 4.17. *Consider the following staircase matrices A and B*

$$A = \begin{bmatrix} \times & \times & \times \\ & & \times \\ & & \times \\ & & \times \end{bmatrix} \quad \textit{and} \quad B = \begin{bmatrix} \times & \times & \times & & \\ & & \times & \times & \times \end{bmatrix}.$$

A and B have both 6 nonzero entries, A has 2 flights, B has 3 flights, and $\operatorname{rank} A = \operatorname{rank} B = 2$.

Next consider the staircase matrices

$$C = \begin{bmatrix} \times & & & & \\ \times & \times & & & \\ & \times & \times & & \\ & & \times & & \\ & & \times & \times & \times \end{bmatrix} \quad \textit{and} \quad D = \begin{bmatrix} \times & & & & \\ \times & \times & \times & & \\ & & \times & & \\ & & \times & \times & \times \\ & & & & \times \end{bmatrix}.$$

C and D have both 9 nonzero entries, C has 6 flights, and D has 5 flights. In addition, $\operatorname{rank} C = 4$ *and* $\operatorname{rank} D = 5$*, that is, the matrix with less flights have larger rank.*

Next, we bound the rank of a generalized staircase matrix B. Since B is constructed by turning some nonzero entries of a staircase matrix A into zero entries, it seems that $\operatorname{rank} B$ has to be smaller than or equal to $\operatorname{rank} A$. This is true, as we will see in Theorem 4.19, but a rigorous proof of this fact requires some work, since for general matrices the operation of turning a nonzero entry into zero may increase the rank. For instance, in MATLAB notation, $\operatorname{rank}[1,1;1,1] = 1$ and $\operatorname{rank}[1,0;1,1] = 2$. The proof of Theorem 4.19 relies on Lemma 4.18.

Lemma 4.18. *If $B \in \mathbb{C}^{m\times p}$ is a generalized staircase matrix, $1 \le i_1 < i_2 < \cdots < i_d \le m$, and $1 \le j_1 < j_2 < \cdots < j_d \le p$, then*

$$\det B(\{i_1, i_2, \ldots, i_d\}, \{j_1, j_2, \ldots, j_d\}) = b_{i_1,j_1}\, b_{i_2,j_2} \cdots b_{i_d,j_d},$$

where $B(\{i_1, i_2, \ldots, i_d\}, \{j_1, j_2, \ldots, j_d\})$ is the submatrix of B that lies in the rows indexed by $\{i_1, i_2, \ldots, i_d\}$ and in the columns indexed by $\{j_1, j_2, \ldots, j_d\}$.

Proof. The proof is by induction on d. For $d = 1$ the result is trivial. Let us assume that the result is true for $d - 1 \ge 1$, and let us prove it for d. Consider that the matrix B is constructed from a staircase matrix A such that (1) $a_{i_1,k} \neq 0$, if $c_1 \le k \le c_1'$, and (2) $a_{i_1,k} = 0$, if $1 \le k \le c_1 - 1$ or $c_1' + 1 \le k \le p$. We split the proof in two cases.

Case 1: $1 \le j_1 \le c_1' - 1$. In this case the definition of generalized staircase matrix implies that all entries in the column j_1 of B below the row i_1 are equal to zero. Then the Laplace expansion of $\det B(\{i_1, i_2, \ldots, i_d\}, \{j_1, j_2, \ldots, j_d\})$ along the first column gives

$$\begin{aligned}\det B(\{i_1, i_2, \ldots, i_d\}, \{j_1, j_2, \ldots, j_d\}) &= b_{i_1,j_1} \det B(\{i_2, \ldots, i_d\}, \{j_2, \ldots, j_d\}) = \\ &= b_{i_1,j_1}\, b_{i_2,j_2} \cdots b_{i_d,j_d},\end{aligned} \tag{4.12}$$

where the last equality follows from the induction hypothesis.

Case 2: $c_1' \le j_1 \le p$. In this case the definition of generalized staircase matrix implies that all entries in the row i_1 of B to the right of the column j_1 are equal to zero. Then the Laplace expansion of $\det B(\{i_1, i_2, \ldots, i_d\}, \{j_1, j_2, \ldots, j_d\})$ along the first row gives again (4.12). □

Theorem 4.19. *Let $A \in \mathbb{C}^{m\times p}$ be a staircase matrix, let $\mathcal{F}(A) = (f_1, f_2, \ldots, f_t)$ be the flight-length sequence of A, and let $\mathcal{L}(A) = (l_1, l_2, \ldots, l_q)$ be the rank-determining list of A. If $B \in \mathbb{C}^{m\times p}$ is any generalized staircase matrix that is obtained by turning some nonzero entries of A into zero entries, then*

$$\operatorname{rank} B \le t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil. \tag{4.13}$$

Proof. Lemma 4.18 and the definition of generalized staircase matrix imply that if a minor of B is nonzero, then the same minor of A is nonzero. So $\operatorname{rank} B \le \operatorname{rank} A$ and the result follows from Theorem 4.16. □

Recall that we used `Algorithm 3` in Theorem 4.12 to construct a staircase matrix $\widetilde{V}_\sigma$ via a bijection σ. It is clear that if we allow zero numbers among the inputs $x_0, x_1, \ldots, x_{n-1}$, then `Algorithm 3` constructs a generalized staircase matrix coming from turning some nonzero entries of $\widetilde{V}_\sigma$ into zero. In addition, according to Definition 4.13 and the discussion in the paragraph just after it, $\mathcal{L}(\sigma) = \mathcal{L}(\widetilde{V}_\sigma)$. Therefore, Corollary 4.20 follows immediately from Theorem 4.19. Here we associate to a bijection σ the magnitude r_σ that will be often used in Section 4.2.

Corollary 4.20. *Let $x_0, x_1, \ldots, x_{n-1}$ be $n \ge 2$ complex numbers not necessarily different from zero, and let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection. Let $\mathcal{L}(\sigma) = (l_1, l_2, \ldots, l_q)$ be the rank-determining list of σ introduced in Definition* 4.13*, and let $\widetilde{V}_\sigma$ be the matrix constructed by* `Algorithm 3`*. Let t be the number of entries of* RCISS(σ)*. Then*

$$\operatorname{rank} \widetilde{V}_\sigma \le r_\sigma, \quad \text{where} \quad r_\sigma := t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil. \tag{4.14}$$

Moreover, if $x_k \neq 0$ for all $k = 0, 1, \ldots, n-1$, then $\operatorname{rank} \widetilde{V}_\sigma = r_\sigma$.

4.1.1 Maximal rank of staircase matrices with a fixed number of nonzero entries

Theorem 4.16 provides a formula for the rank of a staircase matrix A depending on the number of flights and the rank-determining list of A. In this section, once the number of nonzero entries if fixed, we consider the problem of identifying those staircase matrices that have maximal rank. This problem is solved in Theorem 4.22. To get this result, we first prove Lemma 4.21, where we give an upper bound for the rank depending only on the number of nonzero entries, and we provide a necessary condition and a (different) sufficient condition for this bound to be attained. For a given real number x we use the standard notation $\lfloor x \rfloor$ to denote the largest integer which is smaller than or equal to x.

Lemma 4.21. *Let A be a staircase matrix with $n \geq 2$ nonzero entries, and let $\mathcal{F}(A) = (f_1, \ldots, f_t)$ be the flight-length sequence of A. Then*

(a) $\operatorname{rank} A \leq \left\lfloor \dfrac{n+1}{2} \right\rfloor$.

(b) *If $f_i = 1$ for all $i = 1, \ldots, t$, then* $\operatorname{rank} A = \left\lfloor \dfrac{n+1}{2} \right\rfloor$.

(c) *If* $\operatorname{rank} A = \left\lfloor \dfrac{n+1}{2} \right\rfloor$, *then $f_i \leq 2$ for all $i = 1, \ldots, t$.*

Proof. (a) Let $d_c(A)$ and $d_r(A)$ denote the number of columns and rows of A, respectively. Since A is a staircase matrix, we have $d_c(A) + d_r(A) = n + 1$ (this follows easily from `Algorithm 3`). Hence, the result follows from the inequalities $\operatorname{rank} A \leq \min\{d_c(A), d_r(A)\} \leq \frac{n+1}{2}$.

(b) Following the notation of Theorem 4.16, for a staircase matrix A in the conditions of the statement we have $t = n - 1$ and $\mathcal{L}(A) = (n-3)$, so (4.4) gives $\operatorname{rank} A = n - 1 - \left\lceil \frac{n-3}{2} \right\rceil = \left\lfloor \frac{n+1}{2} \right\rfloor$.

(c) We proceed by contradiction. Let $1 \leq i_0 \leq t$ be such that $f_{i_0} \geq 3$. We will construct a staircase matrix $\widehat{A}$ with exactly n nonzero entries and with $\operatorname{rank} \widehat{A} = \operatorname{rank} A + 1$. Using (a) this immediately implies that $\operatorname{rank} A < \left\lfloor \frac{n+1}{2} \right\rfloor$, which contradicts the hypothesis. Let $\widehat{A}$ be a staircase matrix such that

$$\mathcal{F}(\widehat{A}) = (f_1, \ldots, f_{i_0-1}, s_{i_0}, u_{i_0}, v_{i_0}, f_{i_0+1}, \ldots, f_t),$$

where $u_{i_0} = 1$ and s_{i_0}, v_{i_0} are positive integers such that $s_{i_0} + u_{i_0} + v_{i_0} = f_{i_0}$, and $\widehat{A}$ is constructed by creating 3 flights from the i_0th flight of A. This matrix $\widehat{A}$ always exists, since $f_{i_0} \geq 3$. It is obvious that $\widehat{A}$ has n nonzero entries. Now, let us prove that $\operatorname{rank} \widehat{A} = \operatorname{rank} A + 1$. For this, we assume without loss of generality that the i_0th flight of A has all its entries in the same row, we use Gaussian elimination by rows and columns starting from the $(1,1)$ entry on A and $\widehat{A}$, and consider the following two cases:

- If the first (leftmost) entry of the i_0th flight of A (equivalently of $\widehat{A}$) is a pivot, then the i_0th, the (i_0+1)th, and the (i_0+2)th flights of $\widehat{A}$ follow the pattern

$$\begin{array}{cccccccc} \boxed{\times} & \cdots & \times & \times & & & & \\ & & & \boxed{\times} & \times & \cdots & \times & , \end{array}$$

  where $\boxed{\times}$ denote pivot entries. The remaining flights of $\widehat{A}$ have exactly the same structure as the flights of A (all but the i_0th one). As a consequence, $\widehat{A}$ has one more pivot than A.

- If the first (leftmost) entry of the i_0th flight of A is not a pivot, then the i_0th, the (i_0+1)th, and the (i_0+2)th flights of $\widehat{A}$ follow the pattern

$$\begin{array}{ccccccccc} \times & \boxed{\times} & \cdots & \times & & & & & \\ & & & \boxed{\times} & \times & \cdots & \times & , & \end{array}$$

  if $s_{i_0} \geq 2$, or

$$\begin{array}{cccccccc} \times & \boxed{\times} & & & & & & \\ & \times & \boxed{\times} & \times & \cdots & \times & , & \end{array}$$

  if $s_{i_0} = 1$. Again, the remaining flights of $\widehat{A}$ have the same structure as the ones of A (all but the i_0th one), so $\widehat{A}$ has one more pivot than A.

□

Part (b) of Lemma 4.21 provides a particular type of staircase matrices where the maximum rank, given in part (a), is attained. This type corresponds to staircase matrices having only flights of length 1. It is natural to ask whether or not there are other staircase matrices for which this maximum rank is attained. The answer is given in Theorem 4.22, where we provide necessary and sufficient conditions for a staircase matrix A to be of maximal rank, and we prove that this may happen for matrices with flights of lengths larger than 1.

Theorem 4.22. *Let A be a staircase matrix with $n \geq 2$ nonzero entries. Let $\mathcal{F}(A) = (f_1, \ldots, f_t)$ be the flight-length sequence of A, and let $\mathcal{L}(A) = (l_1, \ldots, l_q)$ be the rank-determining list of A. Let α be the number of ones in $\{f_1, f_t\}$. Then $\operatorname{rank} A = \left\lfloor \frac{n+1}{2} \right\rfloor$ if and only if $f_i \leq 2$ for all $i = 1, \ldots, t$ and one of the following sets of conditions hold:*

(a) *n is odd, $\alpha = 2$, and l_i is even for all $i = 1, \ldots, q$;*

(b1) *n is even, $\alpha = 1$, and l_i is even for all $i = 1, \ldots, q$; or*

(b2) *n is even, $\alpha = 2$, and there is exactly one odd element among the elements of $\mathcal{L}(A)$.*

Proof. We will assume from the beginning that $f_i \leq 2$ for all i as a consequence of Lemma 4.21-(c). With this assumption, set n_1 (resp. n_2) for the number of flights of length 1 (resp. 2) of A. Then, following the notation in the statement, we have

$$n_1 = \sum_{i=1}^{q} l_i + \alpha\,, \qquad t = n_1 + n_2\,, \qquad \text{and} \qquad n = n_1 + 2n_2 + 1\,.$$

Hence,

$$n = 2t - \left(\sum_{i=1}^{q} l_i + \alpha \right) + 1. \tag{4.15}$$

Now, we distinguish the cases n odd and n even.

(a) Let n be odd. Then $\operatorname{rank} A = \lfloor (n+1)/2 \rfloor = (n+1)/2$ if and only if

$$\frac{n+1}{2} = t - \sum_{i=1}^{q} \left\lceil \frac{l_i}{2} \right\rceil,$$

by Theorem 4.16. By (4.15), this is equivalent to

$$\sum_{i=1}^{q} \frac{l_i}{2} = \sum_{i=1}^{q} \left\lceil \frac{l_i}{2} \right\rceil + 1 - \frac{\alpha}{2}. \tag{4.16}$$

If $\alpha = 0$ or $\alpha = 1$, then (4.16) is not possible, since

$$\sum_{i=1}^{q} \frac{l_i}{2} < \sum_{i=1}^{q} \left\lceil \frac{l_i}{2} \right\rceil + \frac{1}{2}.$$

Then $\alpha = 2$, and (4.16) holds if and only if l_i is even for all $i = 1, \ldots, q$.

(b) Let n be even. Then $\operatorname{rank} A = \lfloor (n+1)/2 \rfloor = n/2$ if and only if

$$\frac{n}{2} = t - \sum_{i=1}^{q} \left\lceil \frac{l_i}{2} \right\rceil,$$

by Theorem 4.16. Using (4.15) again, this is equivalent to

$$\sum_{i=1}^{q} \frac{l_i}{2} = \sum_{i=1}^{q} \left\lceil \frac{l_i}{2} \right\rceil + \frac{1}{2} - \frac{\alpha}{2}. \tag{4.17}$$

Again, if $\alpha = 0$, then (4.17) does not hold. If $\alpha = 1$, then (4.17) holds if and only if l_i is even for all $i = 1, \ldots, q$. Finally, if $\alpha = 2$, then (4.17) holds if and only if one, and exactly one, among the numbers $l_1, \ldots, l_q$ is odd.

□

Example 4.23 illustrates with staircase matrices having some flights with lengths larger than 1 the three situations presented in Theorem 4.22, where the maximal rank is attained.

Example 4.23. (a) *Let A be the following staircase matrix with* 11 *nonzero entries, which is equivalent to B through elementary row and column replacement operations.*

$$A = \begin{bmatrix} \times & \times & & & & & \\ & \times & & & & & \\ & \times & \times & \times & & & \\ & & & \times & & & \\ & & & \times & \times & \times & \\ & & & & & \times & \end{bmatrix} \sim \begin{bmatrix} \times & 0 & & & & & \\ & \times & & & & & \\ & 0 & \times & 0 & & & \\ & & & \times & & & \\ & & & 0 & \times & 0 & \\ & & & & & \times & \end{bmatrix} = B.$$

The matrix A has $\operatorname{rank} A = 6 = (11+1)/2$, $\mathcal{F}(A) = (1,2,2,2,2,1)$, $\alpha = 2$, *and* $\mathcal{L}(A) = (0)$, *so we are in case* (a) *of Theorem* 4.22.

(b1) *Now, let A be the following staircase matrix with* 12 *nonzero entries*

$$A = \begin{bmatrix} \times & \times & & & & & \\ & \times & & & & & \\ & \times & \times & \times & & & \\ & & & \times & & & \\ & & & \times & \times & \times & \\ & & & & & \times & \\ & & & & & \times & \end{bmatrix} \sim \begin{bmatrix} \times & 0 & & & & & \\ & \times & & & & & \\ & 0 & \times & 0 & & & \\ & & & \times & & & \\ & & & 0 & \times & 0 & \\ & & & & & \times & \\ & & & & & 0 & \end{bmatrix} = B.$$

Then $\operatorname{rank}(A) = 6 = 12/2 = \lfloor (12+1)/2 \rfloor$, $\mathcal{F}(A) = (1,2,2,2,2,2)$, $\alpha = 1$, *and* $\mathcal{L}(A) = (0)$, *so we are in case* (b1) *of Theorem* 4.22.

(b2) *In the last example, the staircase matrix A has 8 nonzero entries.*

$$A = \begin{bmatrix} \times & & & & & \\ \times & \times & \times & & & \\ & & \times & \times & \times & \\ & & & & \times & \end{bmatrix} \sim \begin{bmatrix} \times & & & & & \\ 0 & \times & 0 & & & \\ & & \times & 0 & 0 & \\ & & & & \times & \end{bmatrix} = B.$$

We have rank $A = 4 = 8/2 = \lfloor (8+1)/2 \rfloor$, $\mathcal{F}(A) = (1,2,1,2,1)$, $\alpha = 2$, *and* $\mathcal{L}(A) = (1)$, *so we are in case* (b2) *of Theorem* 4.22.

Notice that if the staircase matrix A is of maximal rank equal to $\lfloor \frac{n+1}{2} \rfloor$, then $\alpha = 0$ cannot occur. The maximum rank $\lfloor \frac{n+1}{2} \rfloor$ considered in Theorem 4.22 is related to Theorem 4.27 in next section. We will explain there this relationship.

4.2 Singular values of Fiedler matrices

Our first result in this section is Theorem 4.24, which proves that any Fiedler matrix M_σ can be written as $U_\sigma + V_\sigma$, where U_σ is a permutation matrix (and so unitary) and V_σ is a matrix such that after removing all its zero rows and columns becomes a staircase matrix. This property will allow us to bound the rank of V_σ via Corollary 4.20.

Theorem 4.24. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and consider the following algorithm:*

`Algorithm 4.` *Given $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ and a bijection σ, the following algorithm constructs a pair of $n \times n$ matrices U_σ and V_σ.*

If σ has a consecution at 0 then

$$U_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \qquad V_0 = \begin{bmatrix} -a_1 & 0 \\ -a_0 - 1 & 0 \end{bmatrix}$$

else

$$U_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \qquad V_0 = \begin{bmatrix} -a_1 & -a_0 - 1 \\ 0 & 0 \end{bmatrix}$$

endif

for $i = 1 : n-2$

if σ has a consecution at i then

$$U_i = \begin{bmatrix} 0 & 1 & 0 \\ U_{i-1}(:,1) & 0 & U_{i-1}(:, 2:i+1) \end{bmatrix}; \qquad V_i = \begin{bmatrix} -a_{i+1} & 0 & 0 \\ V_{i-1}(:,1) & 0 & V_{i-1}(:, 2:i+1) \end{bmatrix}$$

else

$$U_i = \begin{bmatrix} 0 & U_{i-1}(1,:) \\ 1 & 0 \\ 0 & U_{i-1}(2:i+1,:) \end{bmatrix}; \qquad V_i = \begin{bmatrix} -a_{i+1} & V_{i-1}(1,:) \\ 0 & 0 \\ 0 & V_{i-1}(2:i+1,:) \end{bmatrix}$$

endif

endfor

$U_\sigma = U_{n-2}$

$V_\sigma = V_{n-2}$

Then the following statements hold.

(a) *If M_σ is the Fiedler matrix of $p(z)$ associated with σ, then,*

$$M_\sigma = U_\sigma + V_\sigma \,. \tag{4.18}$$

(b) *U_σ is a permutation matrix and, therefore, it is a unitary matrix.*

(c) *If all the zero rows and columns of* V_σ *are removed, then the resulting matrix is the generalized staircase matrix* $\widetilde{V}_\sigma$ *constructed by* `Algorithm 3` *for the inputs* $x_0 = -a_0 - 1, x_1 = -a_1, \ldots, x_{n-1} = -a_{n-1}$ *and* σ.

(d) *Let* $\mathcal{L}(\sigma) = (l_1, l_2, \ldots, l_q)$ *be the rank-determining list of* σ *introduced in Definition* 4.13*, and let* t *be the number of entries of* RCISS(σ)*. Then*

$$\operatorname{rank} V_\sigma \leq r_\sigma \leq \left\lfloor \frac{n+1}{2} \right\rfloor, \quad \textit{where} \quad r_\sigma := t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil. \tag{4.19}$$

Moreover, if $a_0 + 1 \neq 0$ *and* $a_i \neq 0$ *for all* $i = 1, \ldots, n-1$*, then* $\operatorname{rank} V_\sigma = r_\sigma$.

Proof. Part (a). If we compare `Algorithms 1` in Theorem 2.16 and `4`, then we see that $W_0 = U_0 + V_0$. The proof is an induction on W_i, U_i, and V_i. Assume that $W_{i-1} = U_{i-1} + V_{i-1}$ for some $i - 1 \geq 0$. Then the structures of `Algorithms 1` and `4` make obvious that $W_i = U_i + V_i$. The result follows by taking $i = n - 2$.

Part (b). Again the proof is by induction on U_i. By definition U_0 is a 2×2 permutation matrix. Assume that U_{i-1} for some $i - 1 \geq 0$ is a $(i+1) \times (i+1)$ permutation matrix. Then, `Algorithm 4` implies that U_i is a $(i+2) \times (i+2)$ permutation matrix. The result follows by taking $i = n - 2$.

Part (c). We perform an induction on the matrices V_i and $\widetilde{V}_i$ constructed by `Algorithms 4` and `3`, respectively. It is trivial to see that if we remove all zero rows and columns of V_0, then we obtain $\widetilde{V}_0$. Let us assume that the same is true for V_{i-1} and $\widetilde{V}_{i-1}$ for some $i - 1 \geq 0$, and let us prove the result for V_i and $\widetilde{V}_i$. For this purpose note that the first row and the first column of all matrices in the sequence $\{V_0, V_1, \ldots, V_{n-2}\}$ are not identically zero. Therefore, neither the first row nor the first column of V_{i-1} are removed to get $\widetilde{V}_{i-1}$. With this property in mind, it is clear from `Algorithms 4` and `3` that if we remove all zero rows and columns of V_i, then we get $\widetilde{V}_i$. The result follows by taking $i = n - 2$.

Part (d). Since removing zero rows and columns does not change the rank, we get $\operatorname{rank} V_\sigma = \operatorname{rank} \widetilde{V}_\sigma$, and the result is a direct consequence of Corollary 4.20 and Lemma 4.21. □

Parts (a) *and* (d) *of Theorem* 4.24 *imply, in particular, that any Fiedler matrix* M_σ *associated with a bijection* σ *having a low number* (*compared to* n) *of entries in* RCISS(σ) *can be decomposed as a sum of a unitary matrix* U_σ *plus a low-rank matrix* V_σ. The relationship between the rank of V_σ and the number of singular values of M_σ equal to 1 is established in Lemma 4.25, *which is valid for matrices much more general than Fiedler matrices.* In addition, Lemma 4.25 will allow us to reduce the computation of those singular values of M_σ that are not equal to 1, to the computation of the eigenvalues of a matrix whose size may be much smaller than n.

Lemma 4.25. *Let* $A = U + LR \in \mathbb{C}^{n \times n}$*, where* $U \in \mathbb{C}^{n \times n}$ *is a unitary matrix,* $L \in \mathbb{C}^{n \times r}$*, and* $R \in \mathbb{C}^{r \times n}$*. If* $2r < n$*, then* A *has at least* $n - 2r$ *singular values equal to* 1*, and the other* $2r$ *singular values are the square roots of the eigenvalues of the matrix*

$$H = I + \begin{bmatrix} R \\ L^*U \end{bmatrix} \begin{bmatrix} U^*L + R^*L^*L & R^* \end{bmatrix} \in \mathbb{C}^{2r \times 2r}. \tag{4.20}$$

Proof. The singular values of $A = U + LR$ are the square roots of the eigenvalues of A^*A. In the conditions of the statement,

$$\begin{aligned} A^*A = (U + LR)^*(U + LR) &= U^*U + R^*L^*U + U^*LR + R^*L^*LR \\ &= I + \begin{bmatrix} U^*L + R^*L^*L & R^* \end{bmatrix} \begin{bmatrix} R \\ L^*U \end{bmatrix} =: I + \widetilde{L}\widetilde{R}, \end{aligned}$$

where $\widetilde{L} \in \mathbb{C}^{n\times 2r}$ and $\widetilde{R} \in \mathbb{C}^{2r\times n}$. Therefore $\operatorname{rank}(\widetilde{L}\widetilde{R}) \leq 2r$. Now, recall that the eigenvalues of $\widetilde{R}\widetilde{L} \in \mathbb{C}^{2r\times 2r}$, together with an additional $n-2r$ eigenvalues equal to 0, are the eigenvalues of $\widetilde{L}\widetilde{R} \in \mathbb{C}^{n\times n}$ [87, Theorem 1.3.20]. Hence, the eigenvalues of $H = I + \widetilde{R}\widetilde{L} \in \mathbb{C}^{2r\times 2r}$ together with an additional $n-2r$ eigenvalues equal to 1 are the eigenvalues of $A^*A = I + \widetilde{L}\widetilde{R} \in \mathbb{C}^{n\times n}$. These are, precisely, the squares of the singular values of A. □

The application of Lemma 4.25 to a Fiedler matrix M_σ requires to factorize the matrix V_σ in (4.18) as $V_\sigma = L_\sigma R_\sigma$, where $L_\sigma \in \mathbb{C}^{n\times r_\sigma}$, $R_\sigma \in \mathbb{C}^{r_\sigma\times n}$, and r_σ was defined in (4.19). This is done in Lemma 4.26 via `Algorithm 5`. In this algorithm, submatrices like $L(:, 2:1)$ or $R(2:1, :)$ indicate empty matrices.

Lemma 4.26. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let V_σ be the matrix constructed by* `Algorithm 4`*, and let r_σ be the number defined in* (4.19)*. Consider the following algorithm:*

`Algorithm 5`. *Given $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ and a bijection σ, the following algorithm constructs a pair of matrices L_σ and R_σ.*

if σ has a consecution at 0 then

$$L_{-1} = \begin{bmatrix} -a_0 - 1 \end{bmatrix}; \quad R_{-1} = \begin{bmatrix} 1 \end{bmatrix}; \quad L_0 = \begin{bmatrix} -a_1 \\ -a_0 - 1 \end{bmatrix}; \quad R_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

else

$$L_{-1} = \begin{bmatrix} 1 \end{bmatrix}; \quad R_{-1} = \begin{bmatrix} -a_0 - 1 \end{bmatrix}; \quad L_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad R_0 = \begin{bmatrix} -a_1 & -a_0 - 1 \end{bmatrix}$$

endif
for $i = 1 : n-2$
  if σ has an inversion at $i-1$ and a consecution at i then

$$L_i = \begin{bmatrix} -a_{i+1} & 0 \\ -a_i & L_{i-2}(1,:) \\ 0 & 0 \\ 0 & L_{i-2}(2:i,:) \end{bmatrix}; \quad R_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & R_{i-2} \end{bmatrix}$$

  elseif σ has a consecution at $i-1$ and an inversion at i then

$$L_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & L_{i-2} \end{bmatrix}; \quad R_i = \begin{bmatrix} -a_{i+1} & -a_i & 0 & 0 \\ 0 & R_{i-2}(:,1) & 0 & R_{i-2}(:,2:i) \end{bmatrix}$$

  elseif σ has consecutions at $i-1$ and i then

$$L_i = \begin{bmatrix} -a_{i+1} & 0 \\ L_{i-1}(:,1) & L_{i-1}(:,2:\texttt{end}) \end{bmatrix}; \quad R_i = \begin{bmatrix} R_{i-1}(:,1) & 0 & R_{i-1}(:,2:i+1) \end{bmatrix}$$

  elseif σ has inversions at $i-1$ and i then

$$L_i = \begin{bmatrix} L_{i-1}(1,:) \\ 0 \\ L_{i-1}(2:i+1,:) \end{bmatrix}; \quad R_i = \begin{bmatrix} -a_{i+1} & R_{i-1}(1,:) \\ 0 & R_{i-1}(2:\texttt{end},:) \end{bmatrix}$$

  endif
endfor
$L_\sigma = L_{n-2}$
$R_\sigma = R_{n-2}$

Then $V_\sigma = L_\sigma R_\sigma$, with $L_\sigma \in \mathbb{C}^{n\times r_\sigma}$ and $R_\sigma \in \mathbb{C}^{r_\sigma\times n}$. In addition, if $a_0 + 1 \neq 0$ and $a_i \neq 0$ for all $i = 1, \ldots, n-1$, then $\operatorname{rank} V_\sigma = \operatorname{rank} L_\sigma = \operatorname{rank} R_\sigma = r_\sigma$.

Proof. We prove first $V_\sigma = L_\sigma R_\sigma$. To this purpose, let $\{V_0, V_1, \ldots, V_{n-2}\}$ (recall $V_{n-2} = V_\sigma$) be the sequence of matrices constructed by `Algorithm 4`. In addition, we define $V_{-1} := -a_0 - 1$. We will also consider the sequences $\{L_{-1}, L_0, L_1, \ldots, L_{n-2}\}$ and $\{R_{-1}, R_0, R_1, \ldots, R_{n-2}\}$ of matrices

constructed by `Algorithm 5`. The proof consists of proving by induction that $V_i = L_i R_i$ for $i = -1, 0, 1, \ldots, n-2$ (so, the result follows by taking $i = n-2$). It is obvious that $V_{-1} = L_{-1}R_{-1}$ and $V_0 = L_0 R_0$. With a little bit more of effort, it is also straightforward to show via a direct computation that $V_1 = L_1 R_1$ holds. Let us assume that $V_j = L_j R_j$ for all $j = -1, 0, 1, \ldots, i-1$, with $i-1 \geq 1$, and let us prove $V_i = L_i R_i$. In the first place, it follows immediately from `Algorithm 5` and the induction hypothesis that the sizes of L_i and R_i allow us to multiply them. Next, we have two distinguish the four cases that appear in `Algorithm 5`:

(a) If σ has an inversion at $i-1$ and a consecution at i, then `Algorithm 5` implies that

$$L_i R_i = \begin{bmatrix} -a_{i+1} & 0 \\ -a_i & L_{i-2}(1,:) \\ 0 & 0 \\ 0 & L_{i-2}(2:i,:) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & R_{i-2} \end{bmatrix} = \begin{bmatrix} -a_{i+1} & 0 & 0 \\ -a_i & 0 & L_{i-2}(1,:)R_{i-2} \\ 0 & 0 & 0 \\ 0 & 0 & L_{i-2}(2:i,:)R_{i-2} \end{bmatrix}.$$

By the induction hypothesis $L_{i-2}R_{i-2} = V_{i-2}$, so

$$L_i R_i = \begin{bmatrix} -a_{i+1} & 0 & 0 \\ -a_i & 0 & V_{i-2}(1,:) \\ 0 & 0 & 0 \\ 0 & 0 & V_{i-2}(2:i,:) \end{bmatrix}. \tag{4.21}$$

On the other hand, if σ has an inversion at $i-1$ and a consecution at i, then `Algorithm 4` implies

$$V_{i-1} = \begin{bmatrix} -a_i & V_{i-2}(1,:) \\ 0 & 0 \\ 0 & V_{i-2}(2:i,:) \end{bmatrix} \quad \text{and} \quad V_i = \begin{bmatrix} -a_{i+1} & 0 & 0 \\ -a_i & 0 & V_{i-2}(1,:) \\ 0 & 0 & 0 \\ 0 & 0 & V_{i-2}(2:i,:) \end{bmatrix}. \tag{4.22}$$

Therefore, (4.21) and (4.22) imply that $V_i = L_i R_i$.

(b) If σ has a consecution at $i-1$ and an inversion at i, then the proof is similar to the one of Case (a) and is omitted.

(c) If σ has consecutions at $i-1$ and i, then `Algorithm 4` implies that

$$V_i = \begin{bmatrix} -a_{i+1} & 0 & 0 \\ V_{i-1}(:,1) & 0 & V_{i-1}(:,2:i+1) \end{bmatrix}. \tag{4.23}$$

Before completing the proof, it is needed to prove the following auxiliary result: *if σ has a consecution at k, for some $k = 0, 1, \ldots, n-2$, then the matrix R_k constructed by* `Algorithm 5` *satisfies* $R_k(1,:) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$. By definition, $R_0 = [1, 0]$, so the result is true for $k = 0$. We follow by induction. Assume that $R_{k-1}(1,:) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$ if σ has a consecution at $k-1$ for some $k-1 \geq 0$, and let us prove the result for k. If σ has a consecution at k, then we need to consider only two out of the four cases in `Algorithm 5`: (1) σ has an inversion at $k-1$ and a consecution at k; and (2) σ has a consecution at $k-1$ and a consecution at k. In Case (1), $R_k(1,:) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$ by construction. In Case (2), $R_k(1,:) = \begin{bmatrix} R_{k-1}(1,1) & 0 & R_{k-1}(1, 2:k+1) \end{bmatrix}$ and the result follows from the induction assumption.

Next we continue with the proof. If σ has consecutions at $i-1$ and i, then `Algorithm 5` and the auxiliary result imply that

$$L_i R_i = \begin{bmatrix} -a_{i+1} & 0 \\ L_{i-1}(:,1) & L_{i-1}(:, 2:\texttt{end}) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ R_{i-1}(2:\texttt{end}, 1) & 0 & R_{i-1}(2:\texttt{end}, 2:i+1) \end{bmatrix}$$

$$= \begin{bmatrix} -a_{i+1} & 0 & 0 \\ L_{i-1}(:,1) + L_{i-1}(:,2:\texttt{end})R_{i-1}(2:\texttt{end},1) & 0 & L_{i-1}(:,2:\texttt{end})R_{i-1}(2:\texttt{end},2:i+1) \end{bmatrix}$$
$$= \begin{bmatrix} -a_{i+1} & 0 & 0 \\ V_{i-1}(:,1) & 0 & V_{i-1}(:,2:i+1) \end{bmatrix}, \tag{4.24}$$

where the last equality follows from the induction hypothesis $L_{i-1}R_{i-1} = V_{i-1}$ and the auxiliary result, which implies $R_{i-1}(1,:) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$. Equations (4.23) and (4.24) imply $V_i = L_i R_i$.

(d) If σ has inversions at $i-1$ and i, then the proof is similar to the one of Case (c) and is omitted. We only remark that in this case it is needed to prove the following auxiliary result: *if σ has an inversion at k, for some $k = 0, 1, \ldots, n-2$, then the matrix L_k constructed by* `Algorithm 5` *satisfies* $L_k(:,1) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T$.

Next, we prove that if $a_0 + 1 \neq 0$ and $a_i \neq 0$ for all $i = 1, \ldots, n-1$, then $L_\sigma \in \mathbb{C}^{n \times r_\sigma}$, $R_\sigma \in \mathbb{C}^{r_\sigma \times n}$, and $\operatorname{rank} V_\sigma = \operatorname{rank} L_\sigma = \operatorname{rank} R_\sigma = r_\sigma$. It is very easy to see by induction that if $a_0 + 1 \neq 0$ and $a_i \neq 0$ for all $i = 1, \ldots, n-1$, then the structure of `Algorithm 5` implies that, for $i = 0, 1, \ldots, n-2$, all matrices L_i have full column rank and all matrices R_i have full row rank. In particular, $L_\sigma = L_{n-2} \in \mathbb{C}^{n \times r}$ has full column rank and $R_\sigma = R_{n-2} \in \mathbb{C}^{r \times n}$ has full row rank. Since $V_\sigma = L_\sigma R_\sigma$ and $\operatorname{rank} V_\sigma = r_\sigma$ by Theorem 4.24-(d), we get that $r = r_\sigma$ and $\operatorname{rank} L_\sigma = \operatorname{rank} R_\sigma = r_\sigma$.

Finally, observe that the sizes of the matrices $L_\sigma \in \mathbb{C}^{n \times r}$ and $R_\sigma \in \mathbb{C}^{r \times n}$ depend only on σ and n and not on the specific values of the coefficients $a_0, a_1, \ldots, a_{n-1}$ of $p(z)$. Therefore the sizes of L_σ and R_σ are always $L_\sigma \in \mathbb{C}^{n \times r_\sigma}$ and $R_\sigma \in \mathbb{C}^{r_\sigma \times n}$. □

Finally, as a direct corollary of Theorem 4.24, Lemma 4.25, and Lemma 4.26, we state Theorem 4.27, which is our concluding result on singular values of Fiedler matrices. For completeness, we include again in the statement all notions involved.

Theorem 4.27. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let M_σ be the Fiedler matrix of $p(z)$ associated with σ, let $\mathcal{L}(\sigma) = (l_1, l_2, \ldots, l_q)$ be the rank-determining list of σ introduced in Definition* 4.13*, and let t be the number of entries of* RCISS(σ)*. Let us define*

$$r_\sigma := t - \sum_{j=1}^{q} \left\lceil \frac{l_j}{2} \right\rceil,$$

which depends only on σ and not on $p(z)$. If $2r_\sigma < n$, then the following statements hold.

(a) *M_σ has at least $n - 2r_\sigma$ singular values equal to* 1*.*

(b) *The remaining $2r_\sigma$ singular values of M_σ are the square roots of the eigenvalues of the following $2r_\sigma \times 2r_\sigma$ matrix*

$$H_\sigma(p) = I + \begin{bmatrix} R_\sigma \\ L_\sigma^* U_\sigma \end{bmatrix} \begin{bmatrix} U_\sigma^* L_\sigma + R_\sigma^* L_\sigma^* L_\sigma & R_\sigma^* \end{bmatrix} \in \mathbb{C}^{2r_\sigma \times 2r_\sigma}, \tag{4.25}$$

where $U_\sigma \in \mathbb{C}^{n \times n}$ is the permutation matrix constructed by `Algorithm 4`*, and $L_\sigma \in \mathbb{C}^{n \times r_\sigma}$ and $R_\sigma \in \mathbb{C}^{r_\sigma \times n}$ are the matrices constructed by* `Algorithm 5`*.*

Proof. We combine equation (4.18) with $V_\sigma = L_\sigma R_\sigma$, from Lemma 4.26, to obtain $M_\sigma = U_\sigma + L_\sigma R_\sigma$. Then, apply Lemma 4.25 and get the result. □

Note that if the parameter t is small ($t \ll n$), then r_σ is also small, since $r_\sigma \leq t$, which implies that M_σ has many singular values equal to 1 and that the matrix $H_\sigma(p)$ has a small size. Unfortunately, the potential small size of $H_\sigma(p)$ does not allow us to find explicit formulas for its eigenvalues (as it is illustrated in Example 4.29). This is only possible for Frobenius companion matrices because in this case $H_\sigma(p)$ is 2×2. We use in Example 4.28 the approach of Theorem 4.27 to recover the formulas (1.29) of the singular values of Frobenius companion matrices.

From Theorem 4.24-(d), we have $r_\sigma \leq \lfloor (n+1)/2 \rfloor$. In addition, observe that all Fiedler matrices for which $r_\sigma < \lfloor (n+1)/2 \rfloor$, satisfy $2r_\sigma < n$ and, so, have at least one singular value equal to 1. For those Fiedler matrices with $r_\sigma = \lfloor (n+1)/2 \rfloor$ Theorem 4.27 does not apply and they do not have any *guaranteed* singular value equal to 1. These matrices are characterized as those such that the staircase matrix $\widetilde{V}_\sigma$ in Theorem 4.24-(c) satisfies Theorem 4.22 (recall that $\mathcal{L}(\sigma) = \mathcal{L}(\widetilde{V}_\sigma)$ and that the number of entries of RCISS(σ) and $\mathcal{F}(\widetilde{V}_\sigma)$ are equal). In particular, Theorem 4.27 does not apply to some (but not all) of the pentadiagonal Fiedler matrices introduced in (2.7). We will illustrate this fact in Example 4.30.

Example 4.28. We apply here Theorem 4.27 to the first Frobenius companion matrix C_1. From Section 2.2, we know that C_1 corresponds to a bijection μ_1 with only inversions and with RCISS(μ_1) $= (n-1)$. Therefore, in this case, $t = 1$ and $\mathcal{L}(\mu_1) = (0)$, which implies $r_{\mu_1} = 1$ and that C_1 has at least $n-2$ singular values equal to 1. To determine the remaining 2 singular values, we use `Algorithm 4` to construct U_{μ_1} and `Algorithm 5` to construct L_{μ_1} and R_{μ_1} and we get that U_{μ_1} is the first summand in the right-hand side of (4.1), $L_{\mu_1} = [1, 0, \ldots, 0]^T \in \mathbb{C}^{n\times 1}$, and $R_{\mu_1} = [-a_{n-1}, -a_{n-2}, \ldots, -a_1, -a_0 - 1] \in \mathbb{C}^{1\times n}$ (of course, this can be also seen by simple inspection of (4.1)). With these matrices, we easily obtain

$$H_{\mu_1}(p) = \begin{bmatrix} |a_{n-1}|^2 + \cdots + |a_1|^2 + |a_0|^2 + \overline{a}_0 + 1 & |a_{n-1}|^2 + \cdots + |a_1|^2 + |a_0|^2 + a_0 + \overline{a}_0 + 1 \\ -\overline{a}_0 & -\overline{a}_0 \end{bmatrix}.$$

It is immediate to show that the eigenvalues of this matrix are given by (1.29), whose square roots are the 2 remaining singular values of C_1.

Example 4.29. Here, we apply Theorem 4.27 to the Fiedler matrix that is the transpose of the Fiedler matrix F in (2.8). This Fiedler matrix M_σ is associated with a bijection σ having a consecution at 0 an inversions at $1, 2, \ldots, n-2$. Explicitly, this matrix and its decomposition (4.18) are

$$M_\sigma = \begin{bmatrix} -a_{n-1} & \cdots & \cdots & -a_1 & 1 \\ 1 & & & 0 & 0 \\ & \ddots & & \vdots & \vdots \\ & & 1 & 0 & 0 \\ & & & -a_0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & \cdots & 0 & 1 \\ 1 & & & & 0 \\ & \ddots & & & \vdots \\ & & 1 & & \vdots \\ & & & 1 & 0 \end{bmatrix} + \begin{bmatrix} -a_{n-1} & \cdots & -a_1 & 0 \\ & & \vdots & \vdots \\ & & \vdots & \vdots \\ & & \vdots & \vdots \\ & & -a_0 - 1 & 0 \end{bmatrix}. \quad (4.26)$$

In this case CISS(σ) $= (1, n-2)$, RCISS(σ) $= (1, n-2)$, $t = 2$, and $\mathcal{L}(\sigma) = (0)$. Therefore, $r_\sigma = 2$ and M_σ has at least $n-4$ singular values equal to 1. To determine the remaining 4 singular values, we use again `Algorithms 4` and `5` to construct U_σ, L_σ, and R_σ. The matrix U_σ is the first summand of the right-hand side of (4.26), which is the same matrix as in Example 4.28. For L_σ and U_σ, we obtain

$$L_\sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & -a_0 - 1 \end{bmatrix} \in \mathbb{C}^{n\times 2} \quad \text{and} \quad R_\sigma = \begin{bmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \in \mathbb{C}^{2\times n}.$$

With these matrices, we obtain after some algebra

$$H_\sigma(p) = \begin{bmatrix} 1+\sum_{k=1}^{n-1}|a_k|^2 & -\overline{a}_0 a_1(a_0+1) & \sum_{k=1}^{n-1}|a_k|^2 & -a_1 \\ -\overline{a}_1 & -a_0+|a_0+1|^2 & -\overline{a}_1 & 1 \\ 1 & 0 & 1 & 0 \\ \overline{a}_1(\overline{a}_0+1) & -\overline{a}_0|a_0+1|^2 & \overline{a}_1(\overline{a}_0+1) & -\overline{a}_0 \end{bmatrix}.$$

The square roots of the eigenvalues of $H_\sigma(p)$ are the 4 remaining singular values of M_σ. However, it is not easy to obtain (if possible) explicit expressions for them, although we remark the fact that we have reduced an $n \times n$ (with n arbitrary) singular value problem to a 4×4 eigenvalue problem.

Example 4.30. Our last example illustrates that, except in one case, Theorem 4.27 does not apply to the pentadiagonal Fiedler matrices in (2.7) and, so, these matrices do not have, in general, any singular value equal to 1. Since $P_3 = P_1^T$ and $P_4 = P_2^T$, we consider only P_1 and P_2.

From Section 2.2 we have that P_1 is a Fiedler matrix associated with a bijection σ_1 such that $\mathrm{RCISS}(\sigma_1) = (1, 1, \ldots, 1) \in \mathbb{R}^{n-1}$. Therefore $\mathcal{L}(\sigma_1) = (n-3)$, which gives $r_{\sigma_1} = \lfloor (n+1)/2 \rfloor$ both if n is even or odd.

For P_2 a surprise arises. Also, from Section 2.2, we have that P_2 is a Fiedler matrix associated with a bijection σ_2 such that $\mathrm{RCISS}(\sigma_2) = (2, 1, \ldots, 1) \in \mathbb{R}^{n-2}$, which implies $\mathcal{L}(\sigma_2) = (n-4)$. This implies $r_{\sigma_2} = \lfloor (n+1)/2 \rfloor$ if n is even, but $r_{\sigma_2} = (n-1)/2 < \lfloor (n+1)/2 \rfloor$ if n is odd. Therefore if n is odd, the pentadiagonal matrices P_2 and P_4 have, in general, only one singular value equal to 1.

Chapter 5

Adjugate matrix of $zI - M_\sigma$ with M_σ a Fiedler matrix

Given a Fiedler matrix M_σ of a monic polynomial $p(z)$ as in (1.1), the goal of this chapter is to get an explicit expression for the *adjugate matrix* of $zI - M_\sigma$, denoted by $\mathrm{adj}(zI - M_\sigma)$, where the adjugate of a matrix $A \in \mathbb{C}^{n\times n}$ is given in the following definition (See also, for example, [66, Ch. IV §4]).

Definition 5.1. *Let $A \in \mathbb{C}^{n\times n}$, let $\det A_{ij}$ denotes the determinant of the matrix formed by deleting the ith row and jth column of A, and let $B = (b_{ij}) \in \mathbb{C}^{n\times n}$ be the matrix of cofactors of A, where the (i,j) cofactor of A is given by $b_{ij} = (-1)^{i+j} \det A_{ij}$, for $i, j = 1, 2, \ldots, n$. Then, the transpose of the cofactor matrix of A is called the* adjugate *of A and is denoted by* $\mathrm{adj}(A)$.

Remark 5.2. Sometimes the adjugate of a matrix A is called the *classical adjoint*, or simply, the *adjoint* of A. The adjoint of a matrix may also refer to its corresponding adjoint operator, which is its conjugate transpose. For this reason we prefer to use the term *adjugate*.

The main result in this chapter is Theorem 5.3, where we give two different explicit expressions for $\mathrm{adj}(zI - M_\sigma)$. This is a general theoretical result on Fiedler matrices that will be useful in Chapters 8 and 9, where we study eigenvalue condition numbers and pseudospectra of Fiedler matrices, and the backward errors of computed roots of monic polynomials using Fiedler matrices. Notice that the matrix $\mathrm{adj}(zI - M_\sigma)$ is not a constant matrix, but a matrix polynomial in the variable z. When needed, we use the notation $\mathbb{C}^{n\times m}[z]$ for the set of matrix polynomial of size $n \times m$ with complex coefficients.

An explicit expression for the adjugate in the case of the first and second Frobenius companion matrices is already known (see [53, p. 768], [66, Ch. IV §4] or [150]). If $\{p_k(z)\}_{k=0}^{n-1}$ and $\{p_k^{\mathrm{rev}}(z)\}_{k=0}^{n-1}$ denote the Horner shifts, introduced in Definition 2.13, of the polynomial $p(z)$ and of the reversal polynomial $p^{\mathrm{rev}}(z)$ of $p(z)$ (see Definition 2.14), respectively, then

$$\mathrm{adj}(zI - C_2) = \begin{bmatrix} p_0(z) \\ p_1(z) \\ \vdots \\ p_{n-1}(z) \end{bmatrix} \begin{bmatrix} z^{n-1} & \cdots & z & 1 \end{bmatrix} - p(z) \begin{bmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ z & 1 & \ddots & & & \\ \vdots & z & \ddots & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ z^{n-2} & z^{n-1} & \cdots & z & 1 & 0 \end{bmatrix}, \quad (5.1)$$

or, after some algebraic manipulations,

$$\operatorname{adj}(zI-C_2) = \frac{-1}{z}\begin{bmatrix} p_{n-1}^{\mathrm{rev}}(z^{-1}) \\ \vdots \\ p_1^{\mathrm{rev}}(z^{-1}) \\ p_0^{\mathrm{rev}}(z^{-1}) \end{bmatrix}\begin{bmatrix} z^{n-1} & \cdots & z & 1 \end{bmatrix} + p(z)\begin{bmatrix} z^{-1} & z^{-2} & \cdots & \cdots & z^{-n} \\ & z^{-1} & z^{-2} & \cdots & z^{-n+1} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & z^{-2} \\ & & & & z^{-1} \end{bmatrix} \tag{5.2}$$

and $\operatorname{adj}(zI - C_1) = (\operatorname{adj}(zI - C_2))^T$.

Equations (5.1) and (5.2) have a very particular structure: they are a sum of a rank-1 matrix plus a matrix whose (i, j) entry is of the form $p(z)p_{ij}(z)$, where $p_{ij}(z)$ is a polynomial in z of degree at most $n - 2$ in (5.1), and a polynomial in z^{-1} of degree at most n in (5.2). We will prove in Theorem 5.3 that this structure is shared also by $\operatorname{adj}(zI - M_\sigma)$, for any Fiedler matrix M_σ. The functions $\mathfrak{i}_\sigma$ and $\mathfrak{c}_\sigma$, and the n-tuple EPCIS(σ) (see parts (c) and (d) in Definition 2.8) play an important role in the expressions given in Theorem 5.3 for $\operatorname{adj}(zI - M_\sigma)$.

Theorem 5.3. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $p^{\mathrm{rev}}(z)$ be the reversal polynomial of $p(z)$, and let $p_d(z)$ and $p_d^{\mathrm{rev}}(z)$, for $d = 0, 1, \ldots, n-1$, be the degree d Horner shifts of $p(z)$ and $p^{\mathrm{rev}}(z)$, respectively. Let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection with* $\mathrm{EPCIS}(\sigma) = (v_0, v_1, \ldots, v_{n-1})$*, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. Let $x_\sigma, y_\sigma \in \mathbb{C}^n$ be the vectors whose kth entry is*

$$x_\sigma(k) = \begin{cases} z^{\mathfrak{i}_\sigma(0:n-k-1)}p_{k-1}(z) & \text{if } v_{n-k} = 1, \\ z^{\mathfrak{i}_\sigma(0:n-k-1)} & \text{if } v_{n-k} = 0, \end{cases} \quad \text{and} \quad y_\sigma(k) = \begin{cases} z^{\mathfrak{c}_\sigma(0:n-k-1)}p_{k-1}(z) & \text{if } v_{n-k} = 0, \\ z^{\mathfrak{c}_\sigma(0:n-k-1)} & \text{if } v_{n-k} = 1, \end{cases}$$

for $k = 1, 2, \ldots, n$, and let $A_\sigma \in \mathbb{C}^{n\times n}$ be the matrix whose (i, j) entry is

$$A_\sigma(i,j) = \begin{cases} 0 & \text{if } v_{n-i} = v_{n-j} = 0 \text{ and } i \geq j, \\ z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)} & \text{if } v_{n-i} = v_{n-j} = 0 \text{ and } i < j, \\ z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)} & \text{if } v_{n-i} = v_{n-j} = 1 \text{ and } i > j, \\ 0 & \text{if } v_{n-i} = v_{n-j} = 1 \text{ and } i \leq j, \\ 0 & \text{if } v_{n-i} = 0 \text{ and } v_{n-j} = 1, \\ z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)}p_{j-1}(z) & \text{if } v_{n-i} = 1,\ v_{n-j} = 0 \text{ and } i > j, \\ z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}p_{i-1}(z) & \text{if } v_{n-i} = 1,\ v_{n-j} = 0 \text{ and } i < j, \end{cases}$$

for $i, j = 1, 2, \ldots, n$. Also, let $v_\sigma, w_\sigma \in \mathbb{C}^n$ be the vectors whose kth entry is

$$v_\sigma(k) = \begin{cases} -z^{\mathfrak{i}_\sigma(0:n-k-1)-1}p_{n-k}^{\mathrm{rev}}(z^{-1}) & \text{if } v_{n-k} = 1, \\ z^{\mathfrak{i}_\sigma(0:n-k-1)} & \text{if } v_{n-k} = 0, \end{cases}$$

and

$$w_\sigma(k) = \begin{cases} -z^{\mathfrak{c}_\sigma(0:n-k-1)-1}p_{n-k}^{\mathrm{rev}}(z^{-1}) & \text{if } v_{n-k} = 0, \\ z^{\mathfrak{c}_\sigma(0:n-k-1)} & \text{if } v_{n-k} = 1, \end{cases}$$

for $k = 1, 2, \ldots, n$, and let $B_\sigma \in \mathbb{C}^{n\times n}$ be the matrix whose (i, j) entry is

$$B_\sigma(i,j) = \begin{cases} z^{\mathfrak{c}_\sigma(n-i:n-j-1)-(i-j+1)} & \text{if } v_{n-i} = v_{n-j} = 0 \text{ and } i \geq j, \\ 0 & \text{if } v_{n-i} = v_{n-j} = 0 \text{ and } i < j, \\ 0 & \text{if } v_{n-i} = v_{n-j} = 1 \text{ and } i > j, \\ z^{\mathfrak{i}_\sigma(n-j:n-i-1)-(j-i+1)} & \text{if } v_{n-i} = v_{n-j} = 1 \text{ and } i \leq j, \\ 0 & \text{if } v_{n-i} = 0 \text{ and } v_{n-j} = 1, \\ -z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)-(i-j+1)}p_{n-i}^{\mathrm{rev}}(z^{-1}) & \text{if } v_{n-i} = 1,\ v_{n-j} = 0 \text{ and } i > j, \\ -z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)-(j-i+1)}p_{n-j}^{\mathrm{rev}}(z^{-1}) & \text{if } v_{n-i} = 1,\ v_{n-j} = 0 \text{ and } i < j, \end{cases}$$

for $i, j = 1, 2, \ldots, n$. Then,

$$\operatorname{adj}(zI - M_\sigma) = x_\sigma y_\sigma^T - p(z)A_\sigma \tag{5.3}$$
$$= v_\sigma w_\sigma^T + p(z)B_\sigma. \tag{5.4}$$

Note that the vectors x_σ, y_σ, v_σ and w_σ, and the matrices A_σ and B_σ depend on the variable z, though we drop it for the ease of notation.

Before proving Theorem 5.3 we state and prove some technical lemmas.

Lemma 5.4. *Let x_σ, y_σ and A_σ be the vectors and the matrix defined in Theorem* 5.3*, respectively. Then, A_σ is the unique $n \times n$ matrix satisfying the following two properties:*

(a) *The entries of A_σ are polynomials in z, and*

(b) *all entries of $x_\sigma y_\sigma^T - p(z)A_\sigma$ are polynomials of degree less than or equal to $n-1$.*

Proof. Throughout this proof we use the following notation:

$$q_k(z) := -a_k z^k - a_{k-1}z^{k-1} - \cdots - a_1 z - a_0 = z^{k+1}p_{n-k-1}(z) - p(z),$$

for $k = 0, 1, \ldots, n-1$. Note that $q_k(z)$ is a polynomial of degree k.

To prove part (a), it suffices to see that the exponents of the powers of z appearing in the entries of A_σ are nonnegative. This is immediate by definition of $\mathfrak{i}_\sigma$ and $\mathfrak{c}_\sigma$ (see part (d) in Definition 2.8). To prove part (b) we need to distinguish several cases.

(1) $v_{n-i} = v_{n-j} = 0$ and $i \geq j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to

$$x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) = z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}p_{j-1}(z)$$

which is a polynomial of degree less than or equal to $n-1$, because, using (2.3),

$$\mathfrak{i}_\sigma(0:n-i-1) + \mathfrak{c}_\sigma(0:n-j-1) + j - 1 = \mathfrak{i}_\sigma(0:n-i-1) - \mathfrak{i}_\sigma(0:n-j-1) + n - 1 \leq n-1.$$

(2) $v_{n-i} = v_{n-j} = 0$ and $i < j$: Using (2.3), the (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) &= z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}p_{j-1}(z) - p(z)z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)} \\ &= z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}(z^{n-j+1}p_{j-1}(z) - p(z)) \\ &= z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}q_{n-j}(z), \end{aligned}$$

which is a polynomial of degree less than $n-1$, because

$$\mathfrak{i}_\sigma(n-j+1:n-i-1) + n - j \leq n - i - 1 < n - 1.$$

(3) $v_{n-i} = v_{n-j} = 1$ and $i > j$: Using (2.3), the (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) &= z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}p_{i-1}(z) - p(z)z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)} \\ &= z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)}(z^{n-i+1}p_{i-1}(z) - p(z)) \\ &= z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)}q_{n-i}(z), \end{aligned}$$

which is a polynomial of degree less than $n-1$, because

$$\mathfrak{c}_\sigma(n-i+1:n-j-1) + n - i \leq n - j - 1 < n - 1.$$

(4) $v_{n-i} = v_{n-j} = 1$ and $i \leq j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to

$$x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) = z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}p_{i-1}(z),$$

which is a polynomial of degree less than or equal to $n-1$, because, using (2.3),

$$\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)+i-1 = \mathfrak{c}_\sigma(0:n-j-1)-\mathfrak{c}_\sigma(0:n-i-1)+n-1 \leq n-1.$$

(5) $v_{n-i} = 0$ and $v_{n-j} = 1$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to

$$x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) = z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}$$

which is a polynomial of degree less than or equal to $n-1$, by (2.4).

(6) $v_{n-i} = 1$, $v_{n-j} = 0$ and $i > j$: Using (2.3), the (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) =& z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}p_{i-1}(z)p_{j-1}(z) - \\ & p(z)z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)}p_{j-1}(z) \\ =& z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)}p_{j-1}(z)(z^{n-i+1}p_{i-1}(z) - p(z)) \\ =& z^{\mathfrak{c}_\sigma(n-i+1:n-j-1)}p_{j-1}(z)q_{n-i}(z), \end{aligned}$$

which is a polynomial of degree less than $n-1$, because

$$\mathfrak{c}_\sigma(n-i+1:n-j-1)+j-1+n-i \leq i-j-1+j-1+n-i = n-2.$$

(7) $v_{n-i} = 1$, $v_{n-j} = 0$ and $i < j$: Using (2.3), the (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) =& z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}p_{i-1}(z)p_{j-1}(z) - \\ & p(z)z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}p_{i-1}(z) \\ =& z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}p_{i-1}(z)(z^{n-j+1}p_{j-1}(z) - p(z)) \\ =& z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}p_{i-1}(z)q_{n-j}(z), \end{aligned}$$

which is a polynomial of degree less than $n-1$, because

$$\mathfrak{i}_\sigma(n-j+1:n-i-1)+i-1+n-j \leq j-i-1+i-1+n-j = n-2.$$

Now, suppose that there is another matrix B, whose entries are polynomials in z, and such that the entries of the matrix $x_\sigma y_\sigma^T - p(z)B$ are polynomials in z of degree less than or equal to $n-1$. Let $W_1 = x_\sigma y_\sigma^T - p(z)A_\sigma$ and let $W_2 = x_\sigma y_\sigma^T - p(z)B$, then, $W_1 - W_2 = p(z)(B - A_\sigma)$ is a matrix whose entries are polynomials of degree less than or equal to $n-1$, but if $A_\sigma \neq B$, then $p(z)(B - A_\sigma)$ has, at least, one entry which is a polynomial of degree greater than or equal to n, hence $A_\sigma = B$. □

Lemma 5.5 is key to prove Theorem 5.3. It allows us to relate $\operatorname{adj}(zI - M_\sigma)$ with the adjugate of an $(n-1)\times(n-1)$ matrix obtained by deflating $zI - M_\sigma$ in a certain way. In the following, a matrix polynomial $P(z) \in \mathbb{C}^{n\times n}[z]$ is said to be *unimodular* if $\det P(z)$ is a nonzero constant. In other words, $P(z)$ has a polynomial inverse.

Lemma 5.5. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\sigma : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be a bijection with $\operatorname{PCIS}(\sigma) = (v_0, v_1, \ldots, v_{n-2})$, let M_σ be the Fiedler matrix of $p(z)$ associated with σ, and define the unimodular matrix polynomials $Q(z), R(z) \in \mathbb{C}^{n\times n}[z]$ as*

$$Q(z) := \begin{bmatrix} 1 & 0 & \\ z & 1 & \\ & & I_{n-2} \end{bmatrix} \quad \text{and} \quad R(z) := \begin{bmatrix} 0 & 1 & \\ -1 & p_1(z) & \\ & & I_{n-2} \end{bmatrix}.$$

Then,

(a) *if σ has a consecution at $n-2$,*

$$Q(z)(zI_n - M_\sigma)R(z) = \begin{bmatrix} 1 & \\ & zI_{n-1} - \widetilde{M}_\rho \end{bmatrix},$$

(b) *if σ has an inversion at $n-2$,*

$$R(z)^T(zI_n - M_\sigma)Q(z)^T = \begin{bmatrix} 1 & \\ & zI_{n-1} - \widetilde{M}_\rho \end{bmatrix},$$

where $\rho : \{0, 1, \ldots, n-2\} \to \{1, \ldots, n-1\}$ is a bijection such that PCIS(ρ) $= (v_0, v_1, \ldots, v_{n-3})$, and $\widetilde{M}_\rho = \widetilde{M}_{\rho^{-1}(1)}\widetilde{M}_{\rho^{-1}(2)} \cdots \widetilde{M}_{\rho^{-1}(n-1)}$, with

$$\widetilde{M}_k = \begin{bmatrix} I_{n-k-2} & & & \\ & -a_k & 1 & \\ & 1 & 0 & \\ & & & I_{k-1} \end{bmatrix}, \qquad \text{for } k = 1, 2, \ldots, n-3,$$

and

$$\widetilde{M}_0 = \begin{bmatrix} I_{n-2} & \\ & -a_0 \end{bmatrix}, \qquad \widetilde{M}_{n-2} = \begin{bmatrix} -p_2(z)+z & 1 & \\ 1 & 0 & \\ & & I_{n-3} \end{bmatrix}.$$

Proof. We prove part (a) because part (b) is similar. So, let us assume that σ has a consecution at $n-2$. Then, using the commutativity relations (2.2), the factors of M_σ can be rearranged until M_{n-1} is adjacent on the right to M_{n-2}, that is, $M_\sigma = XM_{n-2}M_{n-1}Y$, where X, Y are products of M_i matrices, with $i < n-2$. Now, since $Q(z)$ and $R(z)$ commute with M_i, for $i < n-2$, we have

$$\begin{aligned}
&Q(z)(zI_n - M_\sigma)R(z) = zQ(z)R(z) - XQ(z)M_{n-2}M_{n-1}R(z)Y \\
&= \begin{bmatrix} 0 & z & 0 & \\ -z & z^2 + zp_1(z) & 0 & \\ 0 & 0 & z & \\ & & & zI_{n-3} \end{bmatrix} - X \begin{bmatrix} -1 & z & 0 & \\ -z & z^2 - a_{n-2} & 1 & \\ 0 & 1 & 0 & \\ & & & I_{n-3} \end{bmatrix} Y \\
&= \begin{bmatrix} 0 & z & 0 & \\ -z & z^2 & 0 & \\ 0 & 0 & z & \\ & & & zI_{n-3} \end{bmatrix} - X \left(\begin{bmatrix} -1 & z & 0 & \\ -z & z^2 - z & 0 & \\ 0 & 0 & 0 & \\ & & & 0_{n-3} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \\ 0 & -p_2(z)+z & 1 & \\ 0 & 1 & 0 & \\ & & & I_{n-3} \end{bmatrix} \right) Y \\
&= \begin{bmatrix} 1 & & & \\ & z & & \\ & & z & \\ & & & zI_{n-3} \end{bmatrix} - X \begin{bmatrix} 0 & 0 & 0 & \\ 0 & -p_2(z)+z & 1 & \\ 0 & 1 & 0 & \\ & & & I_{n-3} \end{bmatrix} Y \\
&= \begin{bmatrix} 1 & \\ & zI_{n-1} \end{bmatrix} - \begin{bmatrix} 0 & \\ & \widetilde{M}_{\rho^{-1}(1)}\widetilde{M}_{\rho^{-1}(2)} \cdots \widetilde{M}_{\rho^{-1}(n-1)} \end{bmatrix} = \begin{bmatrix} 1 & \\ & zI_{n-1} - \widetilde{M}_\rho \end{bmatrix},
\end{aligned}$$

where we have used that $p_2(z) = zp_1(z) + a_{n-2}$ and the fact that multiplying any matrix of the form $\mathrm{diag}(A, 0_{n-2})$, with $A \in \mathbb{C}^{2\times 2}$, by M_k, for $k = 0, 1, \ldots, n-3$, keeps that matrix unchanged. Finally, notice that the relative positions of the matrices $\widetilde{M}_0, \widetilde{M}_1, \ldots, \widetilde{M}_{n-2}$ in $\widetilde{M}_\rho$ are the same as the relative positions of the matrices $M_0, M_1, \ldots, M_{n-2}$ in M_σ, therefore PCIS(ρ) $= (v_0, v_1, \ldots, v_{n-3})$. □

Remark 5.6. Some important observations about the matrix $\widetilde{M}_\rho$ in Lemma 5.5 are in order:

(a) The matrix $\widetilde{M}_i$, for $i = 0, \ldots, n-3$ is obtained from M_i by removing the first row and the first column.

(b) The matrix $\widetilde{M}_\rho$ can be seen formally as a Fiedler matrix of the polynomial $r(z) := z^{n-1} + \sum_{k=0}^{n-2} b_k z^k$, where $b_{n-2} = p_2(z) - z$ and $b_k = a_k$, for $k = 0, 1, \ldots, n-3$. Notice that $r(z) = p(z)$ for all $z \in \mathbb{C}$. We also want to emphasize that the formal $(n-2)$th coefficient of $r(z)$ is not an scalar, but a polynomial in z.

(c) The Horner shifts of $r(z)$ satisfy: $r_0(z) = p_0(z) = 1$ and $r_k(z) = p_{k+1}(z)$ for $k = 1, 2, \ldots, n-2$.

Now, armed with Lemmas 5.4 and 5.5 we are in the position to prove Theorem 5.3.

Proof. (of **Theorem 5.3**) The proof proceeds by induction in n. For $n = 2$ there are only two Fiedler matrices, namely the first and second Frobenius companion matrices. For these two matrices we have

$$\operatorname{adj}(zI - C_2) = \operatorname{adj}\left(\begin{bmatrix} a_1 + z & -1 \\ a_0 & z \end{bmatrix}\right) = \begin{bmatrix} z & 1 \\ -a_0 & a_1 + z \end{bmatrix} = \begin{bmatrix} 1 \\ p_1(z) \end{bmatrix} \begin{bmatrix} z & 1 \end{bmatrix} - p(z) \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

and

$$\operatorname{adj}(zI - C_1) = \operatorname{adj}\left(\begin{bmatrix} a_1 + z & a_0 \\ -1 & z \end{bmatrix}\right) = \begin{bmatrix} z & -a_0 \\ 1 & a_1 + z \end{bmatrix} = \begin{bmatrix} z \\ 1 \end{bmatrix} \begin{bmatrix} 1 & p_1(z) \end{bmatrix} - p(z) \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

which are the matrices in the statement of Theorem 5.3 with $\operatorname{PCIS}(\sigma) = (1)$ and $\operatorname{PCIS}(\sigma) = (0)$, respectively. Assume that the result is true for Fiedler matrices of size $(n-1) \times (n-1)$. To prove it for size $n \times n$, we have to distinguish two cases, namely, whether σ has a consecution or an inversion at $n-2$. Suppose that σ has a consecution at $n-2$ (the proof when σ has an inversion at $n-2$ is similar and we omit it). Then, by Lemma 5.5, we have that

$$zI_n - M_\sigma = Q(z)^{-1} \begin{bmatrix} 1 & \\ & zI_{n-1} - \widetilde{M}_\rho \end{bmatrix} R(z)^{-1},$$

therefore

$$\begin{aligned} \operatorname{adj}(zI_n - M_\sigma) &= \operatorname{adj}\left(R(z)^{-1}\right) \operatorname{adj}\left(\begin{bmatrix} 1 & \\ & zI_{n-1} - \widetilde{M}_\rho \end{bmatrix}\right) \operatorname{adj}\left(Q(z)^{-1}\right) = \\ &= R(z) \begin{bmatrix} p(z) & \\ & \operatorname{adj}(zI_{n-1} - \widetilde{M}_\rho) \end{bmatrix} Q(z), \end{aligned}$$

where we have used the identities $\operatorname{adj}(AB) = \operatorname{adj}(B)\operatorname{adj}(A)$, $\det R(z) = \det Q(z) = 1$, and $\det(zI_{n-1} - \widetilde{M}_\rho) = p(z)$. By the induction hypothesis

$$\begin{aligned} \operatorname{adj}(zI_n - M_\sigma) &= R(z) \begin{bmatrix} p(z) & \\ & x_\rho y_\rho^T - p(z) A_\rho \end{bmatrix} Q(z) = \\ &= R(z) \begin{bmatrix} 0 \\ x_\rho \end{bmatrix} \begin{bmatrix} 0 & y_\rho^T \end{bmatrix} Q(z) - p(z) R(z) \begin{bmatrix} -1 & \\ & A_\rho \end{bmatrix} Q(z). \end{aligned}$$

Note that in the induction step we may see $\widetilde{M}_\rho$ as a Fiedler matrix associated with $r(z) = z^{n-1} + \sum_{k=0}^{n-2} b_k z^k$, with b_i, for $i = 0, \ldots, n-2$, as in Remark 5.6, part (b). To finish the proof it suffices to prove the following three identities:

$$\text{(i)} \quad x_\sigma = R(z) \begin{bmatrix} 0 \\ x_\rho \end{bmatrix}, \quad \text{(ii)} \quad y_\sigma = Q^T(z) \begin{bmatrix} 0 \\ y_\rho \end{bmatrix}, \quad \text{and} \quad \text{(iii)} \quad A_\sigma = R(z) \begin{bmatrix} -1 & \\ & A_\rho \end{bmatrix} Q(z).$$

(i) From the expressions of PCIS(σ) and PCIS(ρ) we have $\mathfrak{i}_\rho(0:k-1) = \mathfrak{i}_\sigma(0:k-1)$, for $k = 1, 2, \ldots, n-2$. Also we have that the Horner shifts of $\widetilde{M}_\rho$ are $p_0(z), p_2(z), \ldots, p_{n-1}(z)$ (see part (c) in Remark 5.6). These observations imply that $x_\rho(k) = x_\sigma(k+1)$, for $k = 2, 3, \ldots, n-1$ (note that, for the permutation ρ, n must be replaced by $n-1$ in the expressions for x_ρ and y_ρ). Therefore

$$R(z)\begin{bmatrix}0\\ x_\rho\end{bmatrix} = \begin{bmatrix}0 & 1 & \\ -1 & p_1(z) & \\ & & I_{n-2}\end{bmatrix}\begin{bmatrix}0\\ z^{\mathfrak{i}_\rho(0:n-3)}\\ x_\rho(2:n-1)\end{bmatrix} = \begin{bmatrix}z^{\mathfrak{i}_\rho(0:n-3)}\\ z^{\mathfrak{i}_\rho(0:n-3)}p_1(z)\\ x_\rho(2:n-1)\end{bmatrix} = \begin{bmatrix}z^{\mathfrak{i}_\sigma(0:n-2)}p_0(z)\\ z^{\mathfrak{i}_\sigma(0:n-3)}p_1(z)\\ x_\sigma(3:n)\end{bmatrix} = x_\sigma,$$

where we have used, since $v_{n-2} = 1$, that $\mathfrak{i}_\sigma(0:n-3) = \mathfrak{i}_\sigma(0:n-2)$ and $p_0(z) = 1$.

(ii) From the expressions of PCIS(σ) and PCIS(ρ) we have $\mathfrak{c}_\rho(0:k-1) = \mathfrak{c}_\sigma(0:k-1)$, for $k = 1, 2, \ldots, n-2$. We also have that the Horner shifts of $\widetilde{M}_\rho$ are $p_0(z), p_2(z), \ldots, p_{n-1}(z)$ (see part (c) in Remark 5.6). These observations imply that $y_\rho(k) = y_\sigma(k+1)$, for $k = 2, 3, \ldots, n-1$. Therefore

$$Q(z)y_\rho = \begin{bmatrix}1 & z & \\ 0 & 1 & \\ & & I_{n-2}\end{bmatrix}\begin{bmatrix}0\\ z^{\mathfrak{c}_\rho(0:n-3)}\\ y_\rho(2:n-1)\end{bmatrix} = \begin{bmatrix}z^{\mathfrak{c}_\rho(0:n-3)+1}\\ z^{\mathfrak{c}_\rho(0:n-3)}\\ y_\rho(2:n-1)\end{bmatrix} = \begin{bmatrix}z^{\mathfrak{c}_\sigma(0:n-2)}\\ z^{\mathfrak{c}_\sigma(0:n-3)}\\ y_\sigma(3:n)\end{bmatrix} = y_\sigma,$$

where we have used, since $v_{n-2} = 1$, that $\mathfrak{c}_\sigma(0:n-2) = \mathfrak{c}_\sigma(0:n-3) + 1$.

(iii) We prove this using Lemma 5.4. From (i) and (ii) we know that

$$\operatorname{adj}(zI - M_\sigma) = x_\sigma y_\sigma^T - p(z)R(z)\begin{bmatrix}-1 & \\ & A_\rho(z)\end{bmatrix}Q(z).$$

But the entries of $R(z)\operatorname{diag}(-1, A_\rho(z))Q(z)$ are polynomials in z and, moreover, the entries of $\operatorname{adj}(zI - M_\sigma)$ are polynomials of degree less than or equal to $n-1$. Therefore, by the uniqueness proved in Lemma 5.4, we get:

$$R(z)\begin{bmatrix}-1 & \\ & A_\rho(z)\end{bmatrix}Q(z) = A_\sigma.$$

To prove (5.4) we only need to check that the (i,j) entry of the matrix $x_\sigma y_\sigma^T - p(z)A_\sigma$ is equal to the (i,j) entry of the matrix $v_\sigma w_\sigma^T + p(z)B_\sigma$. Throughout the rest of the proof we use the following relations between the Horner shifts of $p(z)$ and $p^{\mathrm{rev}}(z)$, and the polynomials $\{q_k(z)\}_{k=0}^{n-1}$, defined in the proof of Lemma 5.4,

$$p_k(z) = z^{-n+k}(p(z) + q_{n-k-1}(z)) \quad \text{and} \quad z^{-k}q_k(z) = -p_k^{\mathrm{rev}}(z^{-1}) \quad \text{for } k = 0, 1, \ldots, n-1.$$

We have to distinguish the same cases as in the proof of Lemma 5.4.

(1) $v_{n-i} = v_{n-j} = 0$ and $i \geq j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is

$$\begin{aligned}x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) =& z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)}p_{j-1}(z)\\ =& z^{\mathfrak{i}_\sigma(0:n-i-1)+\mathfrak{c}_\sigma(0:n-j-1)-n+j-1}(p(z) + q_{n-j}(z))\\ =& z^{\mathfrak{i}_\sigma(0:n-i-1)}\left(-p_{n-j}^{\mathrm{rev}}(z^{-1})z^{\mathfrak{c}_\sigma(0:n-j-1)-1}\right) +\\ & p(z)z^{\mathfrak{c}_\sigma(n-i:n-j-1)-(i-j+1)},\end{aligned}$$

which is equal to $v_\sigma(i)w_\sigma(j) + p(z)B_\sigma(i,j)$.

(2) $v_{n-i} = v_{n-j} = 0$ and $i < j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) =& z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)}p_{j-1}(z) - p(z)z^{\mathrm{i}_\sigma(n-j+1:n-i-1)} \\ =& z^{\mathrm{i}_\sigma(n-j+1:n-i-1)}q_{n-j}(z) \\ =& z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)-n+j-1}q_{n-j}(z) \\ =& z^{\mathrm{i}_\sigma(0:n-i-1)}\left(-z^{\mathrm{c}_\sigma(0:n-j-1)-1}p_{n-j}^{\mathrm{rev}}(z^{-1})\right), \end{aligned}$$

which is equal to $v_\sigma(i)w_\sigma(j) + p(z)B_\sigma(i,j)$

(3) $v_{n-i} = v_{n-j} = 1$ and $i > j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) =& z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)}p_{i-1}(z) - p(z)z^{\mathrm{c}_\sigma(n-i+1:n-j-1)} \\ =& z^{\mathrm{c}_\sigma(n-i+1:n-j-1)}q_{n-i}(z) \\ =& z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)-n+i-1}q_{n-i}(z) \\ =& \left(-z^{\mathrm{i}_\sigma(0:n-i-1)-1}p_{n-i}^{\mathrm{rev}}(z^{-1})\right)z^{\mathrm{c}_\sigma(0:n-j-1)}, \end{aligned}$$

which is equal to $v_\sigma(i)w_\sigma(j) + p(z)B_\sigma(i,j)$.

(4) $v_{n-i} = v_{n-j} = 1$ and $i \leq j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) =& z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)}p_{i-1}(z) \\ =& z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)-n+i-1}(p(z)+q_{n-i}(z)) \\ =& \left(-z^{\mathrm{i}_\sigma(0:n-i-1)-1}p_{n-i}^{\mathrm{rev}}(z^{-1})\right)z^{\mathrm{c}_\sigma(0:n-j-1)} + \\ & p(z)z^{\mathrm{i}_\sigma(n-j:n-i-1)-(j-i+1)}, \end{aligned}$$

which is equal to $v_\sigma(i)w_\sigma(j) + p(z)B_\sigma(i,j)$.

(5) $v_{n-i} = 0$ and $v_{n-j} = 1$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is

$$x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) = z^{\mathrm{i}_\sigma(0:n-i-1)}z^{\mathrm{c}_\sigma(0:n-j-1)},$$

which is equal to $v_\sigma(i)w_\sigma(j) + p(z)B_\sigma(i,j)$.

(6) $v_{n-i} = 1$, $v_{n-j} = 0$ and $i > j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is

$$\begin{aligned} x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) =& z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)}p_{i-1}(z)p_{j-1}(z) - \\ & p(z)z^{\mathrm{c}_\sigma(n-i+1:n-j-1)}p_{j-1}(z) \\ =& z^{\mathrm{c}_\sigma(n-i+1:n-j-1)}p_{j-1}(z)q_{n-i}(z) \\ =& z^{\mathrm{c}_\sigma(n-i+1:n-j-1)-n+j-1}(p(z)+q_{n-j}(z))q_{n-i}(z) \\ =& \left(-z^{\mathrm{i}_\sigma(0:n-i-1)-1}p_{n-i}^{\mathrm{rev}}(z^{-1})\right)\left(-z^{\mathrm{c}_\sigma(0:n-j-1)-1}p_{n-j}^{\mathrm{rev}}(z^{-1})\right) + \\ & p(z)\left(-z^{\mathrm{c}_\sigma(n-i+1:n-j-1)-(i-j+1)}p_{n-i}^{\mathrm{rev}}(z^{-1})\right), \end{aligned}$$

which is equal to $v_\sigma(i)w_\sigma(j) + p(z)B_\sigma(i,j)$.

(7) $v_{n-i} = 1$, $v_{n-j} = 0$ and $i < j$: The (i,j) entry of $x_\sigma y_\sigma^T - p(z)A_\sigma$ is

$$x_\sigma(i)y_\sigma(j) - p(z)A_\sigma(i,j) = z^{\mathrm{i}_\sigma(0:n-i-1)+\mathrm{c}_\sigma(0:n-j-1)}p_{i-1}(z)p_{j-1}(z) -$$

$$
\begin{aligned}
&p(z)z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}p_{i-1}(z)\\
=&z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)}p_{i-1}(z)q_{n-j}(z)\\
=&z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)-n+i-1}(p(z)+q_{n-i}(z))q_{n-j}(z)\\
=&\left(-z^{\mathfrak{i}_\sigma(0:n-i-1)-1}p^{\mathrm{rev}}_{n-i}(z^{-1})\right)\left(-z^{\mathfrak{c}_\sigma(0:n-j-1)-1}p^{\mathrm{rev}}_{n-j}(z^{-1})\right)+\\
&p(z)\left(-z^{\mathfrak{i}_\sigma(n-j+1:n-i-1)-(j-i+1)}p^{\mathrm{rev}}_{n-j}(z^{-1})\right),
\end{aligned}
$$

which is equal to $v_\sigma(i)w_\sigma(j)+p(z)B_\sigma(i,j)$.

□

We illustrate Theorem 5.3 with some 4×4 examples.

Example 5.7. Let $p(z)=z^4+\sum_{k=0}^{3}a_kz^3$ be a monic polynomial of degree 4, and let $\{p_k(z)\}_{k=0}^3$ and $\{p_k^{\mathrm{rev}}(z)\}_{k=0}^3$ be the Horner shifts of $p(z)$ and $p^{\mathrm{rev}}(z)$, respectively. First, we apply Theorem 5.3 to the second Frobenius companion matrix C_2. From Section 2.2 we know that C_2 is a Fiedler matrix associated with a bijection σ_1 such that $\mathrm{EPCIS}(\sigma_1)=(1,1,1,1)$. Then,

$$
\begin{aligned}
\mathrm{adj}(zI-C_2)&=\begin{bmatrix}p_0(z)\\p_1(z)\\p_2(z)\\p_3(z)\end{bmatrix}\begin{bmatrix}z^3 & z^2 & z & 1\end{bmatrix}-p(z)\begin{bmatrix}0&0&0&0\\1&0&0&0\\z&1&0&0\\z^2&z&1&0\end{bmatrix}\\
&=\begin{bmatrix}-z^{-1}p_3^{\mathrm{rev}}(z^{-1})\\-z^{-1}p_2^{\mathrm{rev}}(z^{-1})\\-z^{-1}p_1^{\mathrm{rev}}(z^{-1})\\-z^{-1}p_0^{\mathrm{rev}}(z^{-1})\end{bmatrix}\begin{bmatrix}z^3 & z^2 & z & 1\end{bmatrix}+p(z)\begin{bmatrix}z^{-1}&z^{-2}&z^{-3}&z^{-4}\\0&z^{-1}&z^{-2}&z^{-3}\\0&0&z^{-1}&z^{-2}\\0&0&0&z^{-1}\end{bmatrix}.
\end{aligned}
$$

Second, we apply Theorem 5.3 to the pentadiagonal Fiedler matrix P_1 in (2.7). From Section 2.2 we know that P_1 is a Fiedler matrix associated with a bijection σ_2 such that $\mathrm{EPCIS}(\sigma_2)=(1,0,1,1)$. Then,

$$
\begin{aligned}
\mathrm{adj}(zI-M_\sigma)&=\begin{bmatrix}zp_0(z)\\zp_1(z)\\1\\p_3(z)\end{bmatrix}\begin{bmatrix}z^2 & z & zp_2(z) & 1\end{bmatrix}-p(z)\begin{bmatrix}0&0&p_0(z)&0\\1&0&p_1(z)&0\\0&0&0&0\\z&1&p_2(z)&0\end{bmatrix}\\
&=\begin{bmatrix}-p_3^{\mathrm{rev}}(z^{-1})\\-p_2^{\mathrm{rev}}(z^{-1})\\1\\-z^{-1}p_0^{\mathrm{rev}}(z^{-1})\end{bmatrix}\begin{bmatrix}z^2 & z & -p_1^{\mathrm{rev}}(z^{-1}) & 1\end{bmatrix}-p(z)\begin{bmatrix}z^{-1}&z^{-2}&-z^{-3}p_1^{\mathrm{rev}}(z^{-1})&z^{-3}\\0&z^{-1}&-z^{-2}p_1^{\mathrm{rev}}(z^{-1})&z^{-2}\\0&0&z^{-1}&0\\0&0&-z^{-2}p_0^{\mathrm{rev}}(z^{-1})&z^{-1}\end{bmatrix}.
\end{aligned}
$$

Finally, we apply Theorem 5.3 to the Fiedler matrix F in (2.8). From Section 2.2 we know that F is a Fiedler matrix associated with a bijection σ_3 such that $\mathrm{EPCIS}(\sigma_3)=(0,1,1,1)$. Then,

$$
\begin{aligned}
\mathrm{adj}(zI-F)&=\begin{bmatrix}zp_0(z)\\zp_1(z)\\zp_2(z)\\1\end{bmatrix}\begin{bmatrix}z^2 & z & 1 & p_3\end{bmatrix}-p(z)\begin{bmatrix}0&0&0&p_0\\1&0&0&p_1\\z&1&0&p_2\\0&0&0&0\end{bmatrix}\\
&=\begin{bmatrix}-p_3^{\mathrm{rev}}(z^{-1})\\-p_2^{\mathrm{rev}}(z^{-1})\\-p_1^{\mathrm{rev}}(z^{-1})\\1\end{bmatrix}\begin{bmatrix}z^2 & z & 1 & -z^{-1}p_0^{\mathrm{rev}}(z^{-1})\end{bmatrix}+p(z)\begin{bmatrix}z^{-1}&z^{-2}&z^{-3}&-z^{-4}p_0^{\mathrm{rev}}(z^{-1})\\0&z^{-1}&z^{-2}&-z^{-3}p_0^{\mathrm{rev}}(z^{-1})\\0&0&z^{-1}&-z^{-2}p_0^{\mathrm{rev}}(z^{-1})\\0&0&0&-z^{-1}p_0^{\mathrm{rev}}(z^{-1})\end{bmatrix}.
\end{aligned}
$$

We finish this chapter with Lemmas 5.8 and 5.9 about the matrices A_σ and B_σ, and the vectors x_σ and y_σ, that will be used in Chapter 8.

Lemma 5.8. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let A_σ, B_σ be the matrices defined in Theorem 5.3. Then, the matrices A_σ, B_σ satisfy:*

(a) *The entries of the matrix A_σ are polynomials in z.*

(b) *The entries of the matrix B_σ are polynomials in z^{-1}.*

Proof. Part (a) is proved in Lemma 5.4. To prove part (b), it suffices to see that the exponents of the powers of z appearing in the entries of B_σ are negative. This is immediate by definition of $\mathfrak{i}_\sigma$ and $\mathfrak{c}_\sigma$, which satisfy that $\mathfrak{i}_\sigma(i, j) \leq j - i - 1$ and $\mathfrak{c}_\sigma(i : j) \leq j - i - 1$. □

Lemma 5.9. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let x_σ and y_σ be the vectors in Theorem 5.3. Then, the vectors x_σ and y_σ have one entry identically equal to one.*

Proof. To prove the result we distinguish two cases: when $M_\sigma = C_1, C_2$ and when $M_\sigma \neq C_1, C_2$. The Frobenius companion matrices C_1 and C_2 are Fiedler matrices associated with bijections σ_1 and σ_2, respectively, such that $\text{EPCIS}(\sigma_1) = (0, 0, \ldots, 0) \in \mathbb{R}^n$ and $\text{EPCIS}(\sigma_2) = (1, 1, \ldots, 1) \in \mathbb{R}^n$ (see Section 2.2). Therefore, $x_{\sigma_1} = y_{\sigma_2} = \begin{bmatrix} z^{n-1} & \cdots & z & 1 \end{bmatrix}^T$ and $y_{\sigma_1} = x_{\sigma_2} = \begin{bmatrix} p_0(z) & p_1(z) & \cdots & p_{n-1}(z) \end{bmatrix}^T$. The result follows by inspection of the entries of these two vectors (notice that $p_0(z) = 1$).

If $M_\sigma \neq C_1, C_2$, suppose that $v_0 = 1$, that is, σ has a consecution at 0 (the case $v_0 = 0$ is similar so will be omitted), and let $t \in \{1, 2, \ldots, n-3\}$ be such that $v_{t+1} = 0$ and $v_j = 1$ for any $j \leq t$ (note that a number t satisfying those conditions always exists when $M_\sigma \neq C_1, C_2$). Then, from the expression of the kth entry of x_σ and y_σ given in Theorem 5.3, we get $x_\sigma(n - t - 1) = z^{\mathfrak{i}_\sigma(0:t)} = 1$ and $y_\sigma(n) = z^{\mathfrak{c}_\sigma(0:-1)} = 1$. □

Chapter 6

New bounds for roots of polynomials based on Fiedler companion matrices

Given a monic polynomial $p(z)$ as in (1.1), and λ a root of $p(z)$, we are interested in getting new upper and lower bounds on $|\lambda|$. The goal of this chapter is to study if matrix norms of Fiedler companion matrices may be used to obtain new and sharper lower and upper bounds on $|\lambda|$. The development of such bounds requires first to know simple expressions for some relevant matrix norms of Fiedler matrices, a subject that has been studied at depth in Chapter 3. With these expressions at hand, we show in Theorem 6.1 that norms of Fiedler matrices produce many new bounds, but we also show in Theorem 6.2 that none of them improve significatively the classical bounds obtained from the Frobenius companion matrices, that is, the bounds in Theorem 1.5. However, to improve the results of Theorems 6.1 and 6.2, and following a different approach, we prove in Theorem 6.3 that if the norms of the inverses of Fiedler matrices are used, then another family of new bounds on $|\lambda|$ is obtained and we show in Theorem 6.6 that some of the bounds in this family improve significatively the bounds coming from the Frobenius companion matrices for certain classes of polynomials.

In this chapter, we indicate explicitly the dependence of a Fiedler matrix M_σ on a certain polynomial $q(z)$ by using the notation $M_\sigma(q)$.

6.1 Lower and upper bounds from norms of Fiedler matrices

Since any Fiedler matrix $M_\sigma(p)$ of $p(z)$ has the roots of $p(z)$ as eigenvalues, the same argument that we used to get (1.32) from the Frobenius companion matrices $C_1(p)$ and $C_2(p)$ allows us to prove

$$\left(\|M_\sigma(p^\sharp)\|\right)^{-1} \leq |\lambda| \leq \|M_\sigma(p)\|\,, \tag{6.1}$$

for any root λ of $p(z)$, for any Fiedler matrix of $p(z)$, and for any submultiplicative matrix norm, and where $p^\sharp(z)$ is the monic reversal polynomial of $p(z)$ (see Definition 2.14). As a consequence, it is natural to try to use (6.1) combined with the 1-, 2-, ∞-, and Frobenius norms for obtaining new simple lower and upper bounds on the absolute values of the roots of $p(z)$. Since there exist $2^{n-1}-2$ Fiedler matrices that are different from the Frobenius companion matrices (see Corollary 2.18), this strategy may expand considerably, with respect Theorem 1.5, the arena in which to look for good bounds of type (1.31). But note that, in order to apply (6.1), we need to know which are the expressions for the 1-, 2-, ∞-, and Frobenius norms of Fiedler matrices. The Frobenius

norms of all Fiedler matrices associated with $p(z)$ are equal (see Corollary 3.5), since all of Fiedler matrices have the same nonzero entries, and therefore no new bounds can be obtained from $\|\cdot\|_F$. In addition, we have seen in Chapter 4 that, except in the case of Frobenius companion matrices, simple expressions for the 2-norm of Fiedler matrices are not available, and it seems very difficult to get them. So, in this context, we only investigate which bounds can be derived using the expressions for the ∞- and the 1- norms of any Fiedler matrix obtained in Chapter 3.

As a direct consequence of (6.1) and the expression for the ∞-norm of a Fiedler matrix in Theorem 3.8, we obtain in Theorem 6.1 the first family of new lower and upper bounds for the absolute values of the roots of monic polynomials presented in this chapter. We use the expression "family of lower/upper bounds" because for each different CISS(σ) we obtain a different couple of lower/upper bounds.

Theorem 6.1. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$, and let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection. If λ is a root of $p(z)$, then*

$$\left(\max \left\{ \frac{1}{|a_0|}, \gamma_{\sigma,0}(p^\sharp), \gamma_{\sigma,1}(p^\sharp), \ldots, \gamma_{\sigma,\ell}(p^\sharp) \right\} \right)^{-1} \leq |\lambda| \leq \max\{|a_0|, \gamma_{\sigma,0}(p), \gamma_{\sigma,1}(p), \ldots, \gamma_{\sigma,\ell}(p)\}, \tag{6.2}$$

where the quantities $\gamma_{\sigma,k}(p)$ and $\gamma_{\sigma,k}(p^\sharp)$, for $k = 0, 1, 2, \ldots, \ell$, are those defined in Theorem 3.8 for $p(z)$ and $p^\sharp(z)$, respectively.

Observe that in the statement of Theorem 6.1 we have not imposed $a_0 \neq 0$, which, strictly speaking, is necessary for obtaining the lower bound in (6.2). However, if $a_0 = 0$, then the lower bound can be taken to be zero and this is consistent with the fact that $p(z)$ has at least one root equal to zero.

Theorem 6.2 is the main result in this section. It proves that the bounds coming from applying (6.2) to all Fiedler matrices (i.e., from $\left(\|M_\sigma(p^\sharp)\|_\infty\right)^{-1} \leq |\lambda| \leq \|M_\sigma(p)\|_\infty$) never improve Cauchy's lower (i.e., $\left(\|C_2(p^\sharp)\|_\infty\right)^{-1}$) and Cauchy's upper (i.e., $\|C_2(p)\|_\infty$) bounds by a factor larger than 2. In this sense, the classical Cauchy's bounds in Theorem 1.5 are optimal, up to a factor 2, among those obtained from (6.2) and, in fact, we will see that they are strictly optimal for a large subclass of Fiedler matrices.

Theorem 6.2. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let* CISS$(\sigma) = (\mathfrak{c}_0, \mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{c}_\ell, \mathfrak{i}_\ell)$ *be the consecution-inversion structure sequence of σ. Let $C_2(p)$ be the second Frobenius companion matrix of $p(z)$ and let $M_\sigma(p)$ be the Fiedler matrix of $p(z)$ associated with σ. Then the following statements hold.*

(a) *If $\mathfrak{c}_0 > 0$, then $\|C_2(p)\|_\infty \leq \|M_\sigma(p)\|_\infty$.*
(This means that Cauchy's upper bound is the sharpest upper bound among those in (6.2) when $\mathfrak{c}_0 > 0$.)

(b) *If $\mathfrak{c}_0 = 0$, then $\|C_2(p)\|_\infty - 1 \leq \|M_\sigma(p)\|_\infty$.*
(This means that Cauchy's upper bound is essentially the sharpest upper bound among those in (6.2) when $\|C_2(p)\|_\infty$ is large.)

(c) *If $\mathfrak{c}_0 = 0$, then $\|C_2(p)\|_\infty/2 \leq \|M_\sigma(p)\|_\infty$.*
(This means that none of the upper bounds in (6.2) improves Cauchy's upper bound by a factor larger than two.)

(d) *If $\mathfrak{c}_0 > 0$, then $\left(\|M_\sigma(p^\sharp)\|_\infty\right)^{-1} \leq \left(\|C_2(p^\sharp)\|_\infty\right)^{-1}$.*
(This means that Cauchy's lower bound is the sharpest lower bound among those in (6.2) when $\mathfrak{c}_0 > 0$.)

(e) *If* $\mathfrak{c}_0 = 0$, *then* $\left(\|M_\sigma(p^\sharp)\|_\infty\right)^{-1} \leq \left(\|C_2(p^\sharp)\|_\infty - 1\right)^{-1}$.
(This means that Cauchy's lower bound is essentially the sharpest lower bound among those in (6.2) *when* $\|C_2(p^\sharp)\|_\infty$ *is large.)*

(f) *If* $\mathfrak{c}_0 = 0$, *then* $\left(\|M_\sigma(p^\sharp)\|_\infty\right)^{-1} \leq 2\,\left(\|C_2(p^\sharp)\|_\infty\right)^{-1}$.
(This means that none of the lower bounds in (6.2) *improves Cauchy's lower bound by a factor larger than two.)*

Proof. In this proof we use the notation introduced in Theorems 3.8 and 6.1. Parts (a), (b), and (c) are consequences of the following three inequalities:

$$\text{if } \mathfrak{c}_0 > 0, \text{ then } \ \gamma_{\sigma,0}(p) \geq \max\{1+|a_1|, 1+|a_2|, \ldots, 1+|a_{s_0}|\}; \tag{6.3}$$

$$\text{if } \mathfrak{c}_0 = 0, \text{ then } 1+\gamma_{\sigma,0}(p) \geq \max\{1+|a_1|, 1+|a_2|, \ldots, 1+|a_{s_0}|\}; \tag{6.4}$$

and, for $k = 1, 2, \ldots, \ell$,

$$\gamma_{\sigma,k}(p) \geq \max\{1+|a_{s_{k-1}+1}|, \ldots, 1+|a_{s_k}|\}. \tag{6.5}$$

Proof of Part (a). From (3.6), (6.3), and (6.5), we get that if $\mathfrak{c}_0 > 0$, then

$$\|M_\sigma(p)\|_\infty = \max\{|a_0|, \gamma_{\sigma,0}(p), \ldots, \gamma_{\sigma,\ell}(p)\} \geq \max\{|a_0|, 1+|a_1|, \ldots, 1+|a_{n-1}|\} = \|C_2(p)\|_\infty .$$

Proof of Part (b). From (3.6), (6.4), and (6.5), we get that if $\mathfrak{c}_0 = 0$, then

$$\begin{aligned} 1+\|M_\sigma(p)\|_\infty &= \max\{1+|a_0|, 1+\gamma_{\sigma,0}(p), \ldots, 1+\gamma_{\sigma,\ell}(p)\} \\ &\geq \max\{|a_0|, 1+\gamma_{\sigma,0}(p), \gamma_{\sigma,1}(p), \ldots, \gamma_{\sigma,\ell}(p)\} \\ &\geq \max\{|a_0|, 1+|a_1|, \ldots, 1+|a_{n-1}|\} = \|C_2(p)\|_\infty. \end{aligned}$$

Proof of Part (c). From (3.6), we have that $1 \leq \|M_\sigma(p)\|_\infty$. Therefore, from (b), $\|C_2(p)\|_\infty \leq \|M_\sigma(p)\|_\infty + 1 \leq 2\,\|M_\sigma(p)\|_\infty$, which is part (c).

Proofs of Parts (d), (e), and (f). Parts (a), (b), and (c) have been proved for any monic polynomial $p(z)$. Therefore, they can be applied to $p^\sharp(z)$ for proving parts (d), (e), and (f). □

Observe that there exist polynomials for which the inequalities in parts (b), (c), (e), and (f) of Theorem 6.2 become as close as equalities as desired. Note also that even in the case $\mathfrak{c}_0 = 0$, it is possible to find sufficient conditions on the coefficients of $p(z)$ that guarantee $\|C_2(p)\|_\infty \leq \|M_\sigma(p)\|_\infty$ for wide classes of polynomials and for all Fiedler matrices, and also to find sufficient conditions that guarantee $\left(\|M_\sigma(p^\sharp)\|_\infty\right)^{-1} \leq \left(\|C_2(p^\sharp)\|_\infty\right)^{-1}$ for wide classes of polynomials. We do not pursue this goal here since the inequalities proved in parts (b), (c), (e), and (f) show very clearly that Cauchy's bounds are essentially always the sharpest ones in the family (6.2).

6.2 Lower and upper bounds from norms of inverses of Fiedler matrices

To improve the results in Section 6.1, we follow another strategy based on the fact that for any invertible matrix X, the eigenvalues of X^{-1} are the reciprocals of the eigenvalues of X. So, if $a_0 \neq 0$, the eigenvalues of $M_\sigma(p^\sharp)^{-1}$ are the roots of $p(z)$, the eigenvalues of $M_\sigma(p)^{-1}$ are the reciprocals of the roots of $p(z)$, and

$$\left(\|M_\sigma(p)^{-1}\|\right)^{-1} \leq |\lambda| \leq \|M_\sigma(p^\sharp)^{-1}\|, \tag{6.6}$$

for any root λ of $p(z)$, for any Fiedler matrix of $p(z)$, and for any submultiplicative matrix norm. The practical use of (6.6) requires to know $\|M_\sigma(p)^{-1}\|$ and $\|M_\sigma(p^\sharp)^{-1}\|$ for the 1-, 2-, ∞-, and Frobenius norms. Note that for the Frobenius companion matrices $C_i(p)$, $i = 1, 2$, (6.6) is exactly the same as (1.32) for the 1-, 2-, ∞-, and Frobenius norms, since it is easy to see[1] that $\|C_i(p^\sharp)\| = \|C_i(p)^{-1}\|$ and $\|C_i(p)\| = \|C_i(p^\sharp)^{-1}\|$, and new bounds are not obtained. However, we will prove that the use of other Fiedler matrices in (6.6) gives new bounds for the roots of polynomials and, *more important, that some of these bounds are much sharper than Cauchy's lower/upper bounds in certain cases.*

As a direct consequence of (6.6) and the expression in Theorem 3.8 for the ∞-norm of the inverse of a Fiedler matrix we obtain in Theorem 6.3 the second family of new lower and upper bounds for the absolute values of the roots of monic polynomials presented in this chapter. The key difference between Theorem 6.3 and Theorem 6.1 is that some of the bounds presented in Theorem 6.3 improve significantly the classical Cauchy's bounds for wide classes of polynomials. To prove this fact is one of the main goals in this section.

Theorem 6.3. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$, let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$, and let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection. If λ is a root of $p(z)$, then*

$$\left(\max\left\{\frac{1}{|a_0|}, \delta_\sigma(p), \gamma_{\sigma,1}(p), \ldots, \gamma_{\sigma,\ell}(p)\right\}\right)^{-1} \leq |\lambda| \leq \max\{|a_0|, \delta_\sigma(p^\sharp), \gamma_{\sigma,1}(p^\sharp), \ldots, \gamma_{\sigma,\ell}(p^\sharp)\}, \tag{6.7}$$

where the quantities $\delta_\sigma(p)$, $\gamma_{\sigma,k}(p)$, for $k = 1, 2, \ldots, \ell$, and $\delta_\sigma(p^\sharp)$, $\gamma_{\sigma,k}(p^\sharp)$, for $k = 1, 2, \ldots, \ell$, are those defined in Theorem 3.8 for $p(z)$ and $p^\sharp(z)$, respectively.

For making comparisons, a key property that the reader should bear in mind is that Cauchy's and Montel's lower and upper bounds in Theorem 1.5 are included among the bounds in (6.7) for appropriate choices of σ. This is a consequence of the more general result presented in Theorem 6.4.

Theorem 6.4. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$, and let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$. Let $C_1(p)$ and $C_2(p)$ be the first and second Frobenius companion matrices of $p(z)$. Then*

$$C_1(p)^{-1} = R\, C_1(p^\sharp)\, R, \qquad \text{and} \qquad C_2(p)^{-1} = R\, C_2(p^\sharp)\, R, \tag{6.8}$$

where R is the reverse identity matrix, i.e.,

$$R = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

As a consequence,

(a) $\|C_i(p)^{-1}\|_s = \|C_i(p^\sharp)\|_s$, *for $i = 1, 2$ and $s = 1, 2, \infty, F$,*

(b) $\|C_i(p^\sharp)^{-1}\|_s = \|C_i(p)\|_s$, *for $i = 1, 2$ and $s = 1, 2, \infty, F$.*

Proof. The equalities in (6.8) follow from direct matrix multiplication, from the fact that the inverses of $C_1(p)$ and $C_2(p)$ are given by

$$C_1(p)^{-1} = \begin{bmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -1/a_0 & -a_{n-1}/a_0 & \cdots & -a_1/a_0 \end{bmatrix} \quad \text{and} \quad C_2(p)^{-1} = \begin{bmatrix} 0 & \ldots & 0 & -1/a_0 \\ 1 & & & -a_{n-1}/a_0 \\ & \ddots & & \vdots \\ & & 1 & -a_1/a_0 \end{bmatrix},$$

[1] These equalities are proved in Theorem 6.4.

and from the expressions of the coefficients of $p^\sharp(z)$. Then, part (a) follows from (6.8) and the fact that 1-, 2-, ∞-, and Frobenius-norms are invariant under multiplication by the matrix R. Finally, part (b) follows from applying part (a) to $p^\sharp$ and the fact that $(p^\sharp)^\sharp = p$. □

Recall that Cauchy's and Montel's upper bounds are $\|C_2(p)\|_\infty$ and $\|C_1(p)\|_\infty$, respectively. So part (b) of Theorem 6.4 allows us to express Cauchy's upper bound as $\|C_2(p^\sharp)^{-1}\|_\infty = \|C_2(p)\|_\infty$ and Montel's upper bound as $\|C_1(p^\sharp)^{-1}\|_\infty = \|C_1(p)\|_\infty$. Since the upper bound in (6.7) is $\|M_\sigma(p^\sharp)^{-1}\|_\infty$, we see that Cauchy's and Montel's upper bounds are included among the upper bounds in (6.7). Analogously, part (a) of Theorem 6.4 allows us to see that Cauchy's lower bound is $\left(\|C_2(p)^{-1}\|_\infty\right)^{-1} = \left(\|C_2(p^\sharp)\|_\infty\right)^{-1}$, and that Montel's lower bound is $\left(\|C_1(p)^{-1}\|_\infty\right)^{-1} = \left(\|C_1(p^\sharp)\|_\infty\right)^{-1}$. Since the lower bound in (6.7) is $\left(\|M_\sigma(p)^{-1}\|_\infty\right)^{-1}$, we see that Cauchy's and Montel's lower bounds are two of the lower bounds in (6.7).

The Fiedler matrix $F(p)$ defined in (2.8) will play a relevant role in determining which are the sharpest bounds among those in (6.7). Recall that the matrix $F(p)$ is associated with any bijection τ such that $\mathrm{CISS}(\tau) = (0, 1, n-2, 0)$ and the explicit expressions of $F(p)$ and $F(p)^{-1}$ are

$$F(p) = \begin{bmatrix} -a_{n-1} & 1 & & & \\ \vdots & & \ddots & & \\ -a_2 & & & 1 & \\ -a_1 & & & & -a_0 \\ 1 & & & & 0 \end{bmatrix} \quad \text{and} \quad F(p)^{-1} = \begin{bmatrix} 0 & & & & 1 \\ 1 & & & & a_{n-1} \\ & \ddots & & & \vdots \\ & & 1 & & a_2 \\ & & & -1/a_0 & -a_1/a_0 \end{bmatrix}. \tag{6.9}$$

The bounds (6.7) for $F(p)$ are summarized in Theorem 6.5 for future reference. These bounds are one of the most important contributions in this chapter, since as it is explained in Theorems 6.7 and 6.9, they improve significantly Cauchy's upper and lower bounds for certain polynomials. Theorem 6.5 follows immediately from (6.6), the expression of $F(p)^{-1}$ in (6.9), and the expression for $F(p^\sharp)^{-1}$ that is obtained from applying the second expression in (6.9) to $p^\sharp(z)$.

Theorem 6.5. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with $n \geq 2$ and $a_0 \neq 0$, let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$, and let $F(p)$ be the Fiedler matrix in* (6.9)*. Then*

(a) $\left(\|F(p)^{-1}\|_\infty\right)^{-1} = \min\left\{\frac{|a_0|}{1+|a_1|}, \frac{1}{1+|a_2|}, \ldots, \frac{1}{1+|a_{n-1}|}\right\}$;

(b) $\|F(p^\sharp)^{-1}\|_\infty = \max\left\{1 + \frac{|a_1|}{|a_0|}, 1 + \frac{|a_2|}{|a_0|}, \ldots, 1 + \frac{|a_{n-2}|}{|a_0|}, |a_0| + |a_{n-1}|\right\}$; *and,*

(c) $\left(\|F(p)^{-1}\|_\infty\right)^{-1} \leq |\lambda| \leq \|F(p^\sharp)^{-1}\|_\infty$, *that is,*

$$\min\left\{\frac{|a_0|}{1+|a_1|}, \frac{1}{1+|a_2|}, \ldots, \frac{1}{1+|a_{n-1}|}\right\} \leq |\lambda| \leq \max\left\{1 + \frac{|a_1|}{|a_0|}, 1 + \frac{|a_2|}{|a_0|}, \ldots, 1 + \frac{|a_{n-2}|}{|a_0|}, |a_0| + |a_{n-1}|\right\}. \tag{6.10}$$

Theorem 6.6 is the first important result on comparison of bounds in this section. It proves that either Cauchy's lower/upper bounds or the lower/upper bounds in part (c) of Theorem 6.5 are essentially the sharpest bounds among those coming from applying (6.7) to all Fiedler matrices. The absolute value of the zero degree coefficient of $p(z)$ is the key to distinguish whether Cauchy's bounds or the ones in Theorem 6.5 are the sharpest. In contrast, $|a_0|$ did not play any role in Theorem 6.2.

Theorem 6.6. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let $\mathrm{CISS}(\sigma) = (\mathfrak{c}_0, \mathfrak{i}_0, \mathfrak{c}_1, \mathfrak{i}_1, \ldots, \mathfrak{c}_\ell, \mathfrak{i}_\ell)$ be the consecution-inversion structure sequence of σ. Let*

$C_2(p)$ be the second Frobenius companion matrix of $p(z)$, let $F(p)$ be the Fiedler matrix of $p(z)$ in (6.9)*, and let $M_\sigma(p)$ be the Fiedler matrix of $p(z)$ associated with σ. Then the following statements hold.*

(a) *If $|a_0| \geq 1$ and $\mathfrak{c}_0 = 0$, then $\|F(p^\sharp)^{-1}\|_\infty \leq \|M_\sigma(p^\sharp)^{-1}\|_\infty$.*
(This means that $F(p)$ gives the sharpest upper bound among the upper bounds in (6.7) *when $|a_0| \geq 1$ and $\mathfrak{c}_0 = 0$.)*

(b) *If $|a_0| \geq 1$ and $\mathfrak{c}_0 > 0$, then $\|F(p^\sharp)^{-1}\|_\infty/2 \leq \|M_\sigma(p^\sharp)^{-1}\|_\infty$.*
(This means that, when $|a_0| \geq 1$ and $\mathfrak{c}_0 > 0$, none of the upper bounds in (6.7) *improves the upper bound given by $F(p)$ by a factor larger than two.)*

(c) *If $|a_0| < 1$ and $\mathfrak{c}_0 > 0$, then $\|C_2(p^\sharp)^{-1}\|_\infty \leq \|M_\sigma(p^\sharp)^{-1}\|_\infty$.*
(This means that Cauchy's upper bound is the sharpest upper bound among those in (6.7) *when $|a_0| < 1$ and $\mathfrak{c}_0 > 0$.)*

(d) *If $|a_0| < 1$ and $\mathfrak{c}_0 = 0$, then $\|C_2(p^\sharp)^{-1}\|_\infty - 1 \leq \|M_\sigma(p^\sharp)^{-1}\|_\infty$.*
(This means that Cauchy's upper bound is essentially the sharpest upper bound among those in (6.7) *when $|a_0| < 1$, $\mathfrak{c}_0 = 0$, and $\|C_2(p^\sharp)^{-1}\|_\infty$ is large.)*

(e) *If $|a_0| < 1$ and $\mathfrak{c}_0 = 0$, then $\|C_2(p^\sharp)^{-1}\|_\infty/2 \leq \|M_\sigma(p^\sharp)^{-1}\|_\infty$.*
(This means that, when $|a_0| < 1$ and $\mathfrak{c}_0 = 0$, none of the upper bounds in (6.7) *improves Cauchy's upper bound by a factor larger than two.)*

(f) *If $|a_0| \leq 1$ and $\mathfrak{c}_0 = 0$, then $\left(\|M_\sigma(p)^{-1}\|_\infty\right)^{-1} \leq \left(\|F(p)^{-1}\|_\infty\right)^{-1}$.*
(This means that $F(p)$ gives the sharpest lower bound among the lower bounds in (6.7) *when $|a_0| \leq 1$ and $\mathfrak{c}_0 = 0$.)*

(g) *If $|a_0| \leq 1$ and $\mathfrak{c}_0 > 0$, then $\left(\|M_\sigma(p)^{-1}\|_\infty\right)^{-1} \leq 2\left(\|F(p)^{-1}\|_\infty\right)^{-1}$.*
(This means that, when $|a_0| \leq 1$ and $\mathfrak{c}_0 > 0$, none of the lower bounds in (6.7) *improves the lower bound given by $F(p)$ by a factor larger than two.)*

(h) *If $|a_0| > 1$ and $\mathfrak{c}_0 > 0$, then $\left(\|M_\sigma(p)^{-1}\|_\infty\right)^{-1} \leq \left(\|C_2(p)^{-1}\|_\infty\right)^{-1}$.*
(This means that Cauchy's lower bound is the sharpest lower bound among those in (6.7) *when $|a_0| > 1$ and $\mathfrak{c}_0 > 0$.)*

(i) *If $|a_0| > 1$ and $\mathfrak{c}_0 = 0$, then $\left(\|M_\sigma(p)^{-1}\|_\infty\right)^{-1} \leq \left(\|C_2(p)^{-1}\|_\infty - 1\right)^{-1}$.*
(This means that Cauchy's lower bound is essentially the sharpest lower bound among those in (6.7) *when $|a_0| > 1$, $\mathfrak{c}_0 = 0$, and $\|C_2(p)^{-1}\|_\infty$ is large.)*

(j) *If $|a_0| > 1$ and $\mathfrak{c}_0 = 0$, then $\left(\|M_\sigma(p)^{-1}\|_\infty\right)^{-1} \leq 2\left(\|C_2(p)^{-1}\|_\infty\right)^{-1}$.*
(This means that, when $|a_0| > 1$ and $\mathfrak{c}_0 = 0$, none of the lower bounds in (6.7) *improves Cauchy's lower bound by a factor larger than two.)*

Proof. The expression for the monic reversal polynomial of $p(z)$ in Definition 2.14 implies that, $p(0)$, i.e., the zero-degree coefficient of $p(z)$, satisfies $|p(0)| = |a_0| \geq 1$ (resp., $|p(0)| = |a_0| < 1$) if and only if $|p^\sharp(0)| = 1/|a_0| \leq 1$ (resp., $|p^\sharp(0)| = 1/|a_0| > 1$). From this, we see: that part (f) applied to $p^\sharp(z)$ implies part (a); that part (g) applied to $p^\sharp(z)$ implies part (b); that part (h) applied to $p^\sharp(z)$ implies part (c); that part (i) applied to $p^\sharp(z)$ implies part (d); and that part (j) applied to $p^\sharp(z)$ implies part (e). Therefore we only need to prove parts (f), (g), (h), (i), and (j). We will use the notation in Theorem 3.8 throughout the proof.

Proof of part (f). If $|a_0| \leq 1$ and $\mathfrak{c}_0 = 0$, then

$$\max\left\{\frac{1}{|a_0|}, \delta_\sigma(p)\right\} = \delta_\sigma(p) \geq \max\left\{\frac{1}{|a_0|} + \frac{|a_1|}{|a_0|}, 1 + |a_2|, \ldots, 1 + |a_{s_0}|\right\}.$$

This inequality, together with (6.5), (3.7), and (6.9) imply $\|M_\sigma(p)^{-1}\|_\infty \geq \|F(p)^{-1}\|_\infty$.

Proof of Part (g). If $|a_0| \leq 1$ and $\mathfrak{c}_0 > 0$, then

$$\max\left\{\frac{1}{|a_0|}, \delta_\sigma(p)\right\} = \max\left\{\frac{1}{|a_0|}, 1+\frac{|a_1|}{|a_0|}, \ldots, 1+\frac{|a_{\mathfrak{c}_0-1}|}{|a_0|}, 1+\frac{|a_{\mathfrak{c}_0}|}{|a_0|}+|a_{\mathfrak{c}_\circ+1}|+\cdots+|a_{s_0}|\right\}$$
$$\geq \max\left\{\frac{1}{|a_0|}, 1+\frac{|a_1|}{|a_0|}, 1+|a_2|, \ldots, 1+|a_{s_0}|\right\} \geq \frac{1}{2}\max\left\{\frac{1}{|a_0|}+\frac{|a_1|}{|a_0|}, 1+|a_2|, \ldots, 1+|a_{s_0}|\right\},$$

where in the first inequality we have used that $|a_0| \leq 1$. In addition, from (6.5), for $k = 1, 2, \ldots, \ell$,

$$\gamma_{\sigma,k}(p) \geq \max\left\{1+|a_{s_{k-1}+1}|, \ldots, 1+|a_{s_k}|\right\} \geq \frac{1}{2}\max\left\{1+|a_{s_{k-1}+1}|, \ldots, 1+|a_{s_k}|\right\}.$$

Combining these results with (3.7) and (6.9), we get $\|M_\sigma(p)^{-1}\|_\infty \geq \frac{1}{2}\|F(p)^{-1}\|_\infty$.

Proof of Part (h). If $|a_0| > 1$ and $\mathfrak{c}_0 > 0$, then

$$\max\left\{\frac{1}{|a_0|}, \delta_\sigma(p)\right\} = \max\left\{\frac{1}{|a_0|}, 1+\frac{|a_1|}{|a_0|}, \ldots, 1+\frac{|a_{\mathfrak{c}_0-1}|}{|a_0|}, 1+\frac{|a_{\mathfrak{c}_0}|}{|a_0|}+|a_{\mathfrak{c}_\circ+1}|+\cdots+|a_{s_0}|\right\}$$
$$\geq \max\left\{\frac{1}{|a_0|}, 1+\frac{|a_1|}{|a_0|}, 1+\frac{|a_2|}{|a_0|}, \ldots, 1+\frac{|a_{s_0}|}{|a_0|}\right\}.$$

In addition, from (6.5), for $k = 1, 2, \ldots, \ell$,

$$\gamma_{\sigma,k}(p) \geq \max\left\{1+|a_{s_{k-1}+1}|, \ldots, 1+|a_{s_k}|\right\} \geq \max\left\{1+\frac{|a_{s_{k-1}+1}|}{|a_0|}, \ldots, 1+\frac{|a_{s_k}|}{|a_0|}\right\}.$$

Combining these results with (3.7) and the expression for $C_2(p)^{-1}$, given in Theorem 6.4, we get $\|M_\sigma(p)^{-1}\|_\infty \geq \|C_2(p)^{-1}\|_\infty$.

Proof of Part (i). If $|a_0| > 1$ and $\mathfrak{c}_0 = 0$, then

$$\max\left\{\frac{1}{|a_0|}, \delta_\sigma(p)\right\} \geq \frac{1}{|a_0|}+\frac{|a_1|}{|a_0|}+\cdots+\frac{|a_{s_0}|}{|a_0|} = \left(\frac{1}{|a_0|}+\frac{|a_1|}{|a_0|}+\cdots+\frac{|a_{s_0}|}{|a_0|}+1\right)-1$$
$$\geq \max\left\{\frac{1}{|a_0|}, 1+\frac{|a_1|}{|a_0|}, 1+\frac{|a_2|}{|a_0|}, \ldots, 1+\frac{|a_{s_0}|}{|a_0|}\right\}-1,$$

and, for $k = 1, 2, \ldots, \ell$,

$$\gamma_{\sigma,k}(p) \geq \max\left\{1+\frac{|a_{s_{k-1}+1}|}{|a_0|}, \ldots, 1+\frac{|a_{s_k}|}{|a_0|}\right\} \geq \max\left\{1+\frac{|a_{s_{k-1}+1}|}{|a_0|}, \ldots, 1+\frac{|a_{s_k}|}{|a_0|}\right\}-1.$$

Combining these results with (3.7) and the expression for $C_2(p)^{-1}$, given in Theorem 6.4, we get $\|M_\sigma(p)^{-1}\|_\infty \geq \|C_2(p)^{-1}\|_\infty - 1$.

Proof of Part (j). From Part (i) and the fact that $1 \leq \|M_\sigma(p)^{-1}\|_\infty$, it follows that $\|C_2(p)^{-1}\|_\infty \leq \|M_\sigma(p)^{-1}\|_\infty + 1 \leq 2\,\|M_\sigma(p)^{-1}\|_\infty$. □

Although parts (a) and (b) of Theorem 6.6 tell us that $\|F(p^\sharp)^{-1}\|_\infty$ is essentially the sharpest upper bound among those in (6.7) when $|a_0| \geq 1$, they do not establish whether or not $\|F(p^\sharp)^{-1}\|_\infty$ improves significantly Cauchy's upper bound. Theorem 6.7 shows that it is possible to construct polynomials for which $\|F(p^\sharp)^{-1}\|_\infty$ can be extremely smaller than Cauchy's upper bound.

Theorem 6.7. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$, let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$, let $U_C(p)$ be the Cauchy's upper bound for $p(z)$, and let $F(p^\sharp)$ be the Fiedler companion matrix of $p^\sharp(z)$ defined in* (6.9)*. If the coefficients of $p(z)$ satisfy*

$$\max\{|a_1|, \ldots, |a_{n-2}|\} \geq |a_0|(|a_0| + |a_{n-1}| - 1) \quad \textit{and} \quad |a_0| > 1, \tag{6.11}$$

then

$$\frac{U_C(p)}{\|F(p^\sharp)^{-1}\|_\infty} \geq \frac{|a_0|}{2}.$$

Proof. If the inequality (6.11) is satisfied, then

$$\|F(p^\sharp)^{-1}\|_\infty = \max\left\{1 + \frac{|a_1|}{|a_0|}, \ldots, 1 + \frac{|a_{n-2}|}{|a_0|}, |a_0| + |a_{n-1}|\right\} = 1 + \frac{1}{|a_0|}\max\{|a_1|, \ldots, |a_{n-2}|\}$$

and $U_C(p) = \max\{1 + |a_1|, \ldots, 1 + |a_{n-1}|\} \geq 1 + \max\{|a_1|, \ldots, |a_{n-2}|\}$. Therefore,

$$\frac{U_C(p)}{\|F(p^\sharp)^{-1}\|_\infty} \geq \frac{1 + \max\{|a_1|, \ldots, |a_{n-2}|\}}{1 + \frac{1}{|a_0|}\max\{|a_1|, \ldots, |a_{n-2}|\}} \geq \frac{|a_0|}{2},$$

where the last inequality is a particular case of the more general inequality $(1+a)/(1+a/b) \geq b/2$, which is valid for any positive numbers $a > 0$ and $b > 0$ such that $1 + a/2 \geq b/2$. Observe that (6.11) guarantees that these conditions are satisfied with $a = \max\{|a_1|, \ldots, |a_{n-2}|\}$ and $b = |a_0|$ (it may help to distinguish the cases $|a_0| > 2$ and $2 \geq |a_0| > 1$). □

Theorem 6.7 states that if (6.11) is satisfied and $|a_0|$ is very large, then Cauchy's upper bound for the absolute values of the roots of a monic polynomial is much larger than the upper bound $\|F(p^\sharp)^{-1}\|_\infty$. Notice that, however, in order for (6.11) to hold when $|a_0|$ is large, there must be another coefficient of $p(z)$ whose absolute value is at least of order $|a_0|^2$. This is the case of the following example that illustrates Theorem 6.7.

Example 6.8. Consider the monic polynomial $p(z) = z^3 + z^2 + 10^{2m}z + 10^m$, for some integer $m > 0$. For this polynomial we have the following upper bounds

$$\begin{aligned}
&|\lambda| \leq \|F(p^\sharp)^{-1}\|_\infty = 1 + 10^m \approx 10^m, \\
&|\lambda| \leq U_C(p) = 1 + 10^{2m} \approx 10^{2m}, && (Cauchy), \\
&|\lambda| \leq \sqrt{2 + 10^{2m} + 10^{4m}} \approx 10^{2m}, && (Carmichael - Mason),
\end{aligned}$$

and $\max\{|\lambda| : \lambda \text{ is a root of } p(z)\} \approx 10^m$. We display also Carmichael-Mason's upper bound just for completeness, since according to the discussion just after Theorem 1.5, Carmichael-Mason's upper bound cannot improve Cauchy's upper bound by a factor larger than $\sqrt{2}$. We observe that the bound $\|F(p^\sharp)^{-1}\|_\infty$ is essentially optimal, while Cauchy's and Carmichael-Mason's upper bounds are extremely larger than $|\lambda|$ if m is large.

Although parts (f) and (g) of Theorem 6.6 tell us that $\left(\|F(p)^{-1}\|_\infty\right)^{-1}$ is essentially the sharpest lower bound among those in (6.7) when $|a_0| \leq 1$, they do not establish whether or not this bound improves significantly Cauchy's lower bound. Theorem 6.9 shows that it is possible to construct polynomials for which $\left(\|F(p)^{-1}\|_\infty\right)^{-1}$ can be extremely larger than Cauchy's lower bound.

Theorem 6.9. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$, let $L_C(p)$ be Cauchy's lower bound for $p(z)$, and let $F(p)$ be the Fiedler matrix defined in* (6.9)*. If the coefficients of $p(z)$ satisfy*

$$\max\{|a_2|, \ldots, |a_{n-1}|\} \geq \frac{1 + |a_1|}{|a_0|}, \quad \textit{and} \quad |a_0| \leq 1, \tag{6.12}$$

then

$$\frac{\left(\|F(p)^{-1}\|_\infty\right)^{-1}}{L_C(p)} \geq \frac{1}{2|a_0|}. \tag{6.13}$$

Proof. Conditions (6.12) and expression (6.9) imply

$$\|F(p)^{-1}\|_\infty = \max\left\{\frac{1+|a_1|}{|a_0|}, 1+|a_2|, \ldots, 1+|a_{n-1}|\right\} = 1+\max\{|a_2|, \ldots, |a_{n-1}|\},$$

and $\max\{|a_2|, \ldots, |a_{n-1}|\} \geq 1$. Also, we have that

$$L_C(p)^{-1} = \max\left\{\frac{1}{|a_0|}, 1+\frac{|a_1|}{|a_0|}, \ldots, 1+\frac{|a_{n-1}|}{|a_0|}\right\} \geq 1+\frac{\max\{|a_2|, \ldots, |a_{n-1}|\}}{|a_0|}.$$

Then

$$\frac{L_C(p)^{-1}}{\|F(p)^{-1}\|_\infty} \geq \frac{1+\max\{|a_2|, \ldots, |a_{n-1}|\}/|a_0|}{1+\max\{|a_2|, \ldots, |a_{n-1}|\}} \geq \frac{1}{2|a_0|}.$$

The last inequality is a particular case of the general inequality $(1+a/b)/(1+a) \geq 1/(2b)$, which is valid for any numbers such that $b > 0$ and $a \geq 1$. □

Theorem 6.9 states that if (6.12) is satisfied and $|a_0|$ is very small, then Cauchy's lower bound for the absolute values of the roots of a monic polynomial is much smaller than the lower bound $\left(\|F(p)^{-1}\|_\infty\right)^{-1}$. Note that in order for (6.12) to hold when $|a_0|$ is small, at least one of the coefficients $a_2, \ldots, a_{n-1}$ must have a large absolute value. This is the case in Example 6.10, which illustrates Theorem 6.9.

Example 6.10. Consider the monic polynomial $p(z) = z^3 + 2\cdot 10^m z^2 + z + 10^{-m}$, with m a positive integer. For this polynomial we have the following lower bounds

$$\begin{aligned}
&|\lambda| \geq \left(\|F(p)^{-1}\|_\infty\right)^{-1} = \frac{1}{1+2\cdot 10^m} \approx 0.5\cdot 10^{-m}, \\
&|\lambda| \geq L_C(p) = \frac{1}{1+2\cdot 10^{2m}} \approx 0.5\cdot 10^{-2m}, && (Cauchy), \\
&|\lambda| \geq \frac{10^{-m}}{\sqrt{2+10^{-2m}+4\cdot 10^{2m}}} \approx 0.5\cdot 10^{-2m}, && (Carmichael-Mason),
\end{aligned}$$

and $\min\{|\lambda| : \lambda \text{ is a root of } p(z)\} \approx 0.7\cdot 10^{-m}$. As in Example 6.8, Carmichael-Mason's bound is displayed for completeness, since according to the discussion right after Theorem 1.5, Carmichael-Mason's lower bound cannot improve Cauchy's lower bound by a factor larger than $\sqrt{2}$. We observe that the bound $\left(\|F(p)^{-1}\|_\infty\right)^{-1}$ is almost optimal, while Cauchy's and Carmichael-Mason's bounds are extremely smaller than $|\lambda|$ if m is large.

6.3 Bounds from Frobenius norms of inverses of Fiedler matrices

As we commented in the Section 6.1, the use of (6.1) with the Frobenius norm makes no sense since all Fiedler matrices of a given monic polynomial have the same Frobenius norm (3.3) and, therefore, we obtain exactly the same bounds as in part 4 of Theorem 1.5 in all cases. However, the use of (6.6) with the Frobenius norm may produce new bounds, since the inverses of all Fiedler matrices of a given monic polynomial *do not have always* the same Frobenius norm. In fact, given $p(z)$, $\|M_\sigma(p)^{-1}\|_F$ depends only on t_σ, i.e., on the number of initial consecutions or

inversions of σ (see Corollary 3.5). In this context, the purpose of this section is to study the bounds $\left(\|M_\sigma(p)^{-1}\|_F\right)^{-1} \leq |\lambda| \leq \|M_\sigma(p^\sharp)^{-1}\|_F$ for the absolute values of the roots λ of a monic polynomial $p(z)$ and to compare them with Cauchy's lower/upper bounds and with the bounds in Theorem 6.5-(c). The main conclusion is that, although the new bounds coming from the Frobenius norm may be sharper in certain situations, the improvements are never significative.

Theorem 6.11 is a direct consequence of (3.4) and (6.6).

Theorem 6.11. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let t_σ be the number of initial consecutions or inversions of σ. If λ is a root of $p(z)$, then $\left(\|M_\sigma(p)^{-1}\|_F\right)^{-1} \leq |\lambda| \leq \|M_\sigma(p^\sharp)^{-1}\|_F$, that is,*

$$\frac{1}{\sqrt{(n-1) + \frac{1+|a_1|^2+\cdots+|a_{t_\sigma}|^2}{|a_0|^2} + |a_{t_\sigma+1}|^2 + \cdots + |a_{n-1}|^2}} \leq |\lambda| \qquad \text{and} \tag{6.14}$$

$$|\lambda| \leq \sqrt{(n-1) + |a_0|^2 + |a_{n-1}|^2 + |a_{n-2}|^2 + \cdots + |a_{n-t_\sigma}|^2 + \frac{|a_{n-t_\sigma-1}|^2 + \cdots + |a_1|^2}{|a_0|^2}}\,. \tag{6.15}$$

Given $p(z)$, the bounds (6.14) and (6.15) depend only on t_σ. On the other hand, the second companion form $C_2(p)$ is a Fiedler matrix that corresponds to the maximum value of t_σ, i.e., $t_\sigma = n-1$, while the matrix $F(p)$ in (6.9) corresponds to the minimum value $t_\sigma = 1$. This allows us to prove Theorem 6.12 directly from (6.14)-(6.15). The reader should recall that the lower and upper bounds of part 4 in Theorem 1.5 are, respectively, $\left(\|C_2(p^\sharp)\|_F\right)^{-1}$ and $\|C_2(p)\|_F$, which are equal, respectively, to $\left(\|C_2(p)^{-1}\|_F\right)^{-1}$ and $\|C_2(p^\sharp)^{-1}\|_F$, by Theorem 6.4.

Theorem 6.12. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $p^\sharp(z)$ be the monic reversal polynomial of $p(z)$, and let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection. Let $C_2(p)$ be the second Frobenius companion form of $p(z)$, let $F(p)$ be the Fiedler matrix defined in* (6.9)*, and let $M_\sigma(p)$ be the Fiedler matrix of $p(z)$ associated with σ. Then the following statements hold.*

(a) *If $|a_0| \geq 1$, then $\|F(p^\sharp)^{-1}\|_F \leq \|M_\sigma(p^\sharp)^{-1}\|_F$.*
(This means that $F(p)$ gives the sharpest upper bound among the upper bounds in (6.15) *when $|a_0| \geq 1$.)*

(b) *If $|a_0| < 1$, then $\|C_2(p^\sharp)^{-1}\|_F \leq \|M_\sigma(p^\sharp)^{-1}\|_F$.*
(This means that the upper bound in part 4 of Theorem 1.5 is the sharpest upper bound among the upper bounds in (6.15) *when $|a_0| < 1$.)*

(c) *If $|a_0| \leq 1$, then $\left(\|M_\sigma(p)^{-1}\|_F\right)^{-1} \leq \left(\|F(p)^{-1}\|_F\right)^{-1}$.*
(This means that $F(p)$ gives the sharpest lower bound among the lower bounds in (6.14) *when $|a_0| \leq 1$.)*

(d) *If $|a_0| > 1$, then $\left(\|M_\sigma(p)^{-1}\|_F\right)^{-1} \leq \left(\|C_2(p)^{-1}\|_F\right)^{-1}$.*
(This means that the lower bound in part 4 of Theorem 1.5 is the sharpest lower bound among the lower bounds in (6.14) *when $|a_0| > 1$.)*

Part (b) in Theorem 6.12 shows us that when $|a_0| < 1$, the upper bounds in (6.15) are of no interest, since all of them are larger than the upper bound in part 4 of Theorem 1.5, which is larger than Carmichael-Mason upper bound, which in turn is larger than Cauchy's upper bound divided by $\sqrt{2}$. Analogously, part (d) in Theorem 6.12 shows us that when $|a_0| > 1$, the lower bounds in (6.14) are of no interest, since all of them are smaller than the lower bound in part 4 of Theorem 1.5.

However, parts (a) and (c) of Theorem 6.12 suggest that the upper bound $\|F(p^\sharp)^{-1}\|_F$ and/or the lower bound $\left(\|F(p)^{-1}\|_F\right)^{-1}$ might improve in certain situations previously known upper/lower bounds for the absolute values of the roots of monic polynomials. In fact, this is true, but Theorem 6.13 shows that these improvements are never larger than a factor $\sqrt{2}$, that is, the improvements are never really significative. This is shown by comparing these bounds with those established in Theorem 6.5.

Theorem 6.13. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ and $a_0 \neq 0$ and let $F(p)$ be the Fiedler matrix of $p(z)$ defined in* (6.9)*. Then*

(a) $\dfrac{1}{\sqrt{2}}\|F(p^\sharp)^{-1}\|_\infty \leq \|F(p^\sharp)^{-1}\|_F$,

(b) $\dfrac{1}{\sqrt{2}}\left(\|F(p)^{-1}\|_F\right)^{-1} \leq \left(\|F(p)^{-1}\|_\infty\right)^{-1}$.

Proof. Part (a) follows from applying part (b) to $p^\sharp(z)$. Therefore we only prove part (b). We have that

$$\frac{\|F^{-1}(p)\|_F}{\|F^{-1}(p)\|_\infty} = \frac{\sqrt{(n-1) + \frac{1}{|a_0|^2} + \frac{|a_1|^2}{|a_0|^2} + |a_2|^2 + \cdots + |a_{n-1}|^2}}{\max\left\{\frac{1}{|a_0|} + \frac{|a_1|}{|a_0|}, 1 + |a_2|, 1 + |a_3|, \ldots, 1 + |a_{n-1}|\right\}}.$$

Next, use

$$\frac{1}{|a_0|} + \frac{|a_1|}{|a_0|} \leq \sqrt{2}\sqrt{\left(\frac{1}{|a_0|}\right)^2 + \left(\frac{|a_1|}{|a_0|}\right)^2} \quad \text{and} \quad 1 + |a_i| \leq \sqrt{2}\sqrt{1 + |a_i|^2},\ i = 2, \ldots, n-1,$$

and the result follows immediately. □

6.4 Optimal bounds based on norms of diagonal similarities

The bounds in Theorem 1.5 and the ones that can be obtained from Fiedler matrices and their inverses with the 1-, ∞-, and Frobenius norms (see, for instance, (6.10)) have an important drawback: the lower bounds are always smaller than 1 and the upper bounds are always larger than 1. This is a consequence of the presence of entries equal to 1 in any Fiedler matrix and its inverse. For $C_1(p)$ and $C_2(p)$ an standard way to overcome this drawback is to use diagonal similarities, which do not change neither the eigenvalues nor the zero pattern, and to use (1.32). More precisely, let D and $\widetilde{D}$ be nonsingular diagonal matrices, then from (1.32) we get $\left(\|\widetilde{D}^{-1}C_i(p^\sharp)\widetilde{D}\|\right)^{-1} \leq |\lambda| \leq \|D^{-1}C_i(p)D\|$, for $i = 1, 2$. Given a polynomial $p(z)$, the selection of a proper D (and/or $\widetilde{D}$) may improve drastically the bounds, but a choice of D that is good for certain polynomials may be bad for others, so the choice of proper diagonal similarities is not immediate. Some specific D's have been used to get the well-know Fujiwara's [65] and Kojima's bounds [97] (see also [87, p. 319]). The use of diagonal similarities is also possible with Fiedler matrices, both combined with (6.1) and (6.6), and it is possible to obtain explicit expressions of the norms of the matrices involved in these bounds for the 1-, ∞-, and Frobenius norms. However, how to select proper diagonal matrices that improve the known bounds for wide classes of polynomials is not clear. This problem requires further and extensive investigation and in this work we limit ourselves to give some theoretical results on the optimal bounds that can be obtained with this approach. In this context, it should be noted that the Fiedler matrix $F(p)$ is a very particular diagonal similarity of $C_2(p)$ if $a_0 \neq 0$ (both matrices are also similar if $a_0 = 0$, but then the similarity is not diagonal). In fact, $F(p)$ is the only Fiedler matrix of $p(z)$ that is diagonally similar to $C_2(p)$, because other Fiedler matrices have a different zero pattern.

All upper bounds presented in this chapter for the absolute values of the roots λ of $p(z)$, and the majority of the bounds existing in the literature, are functions only of the absolute values of the coefficients of $p(z)$. A well-known bound of this type is the unique positive real root of $u(z) = z^n - \sum_{k=0}^{n-1} |a_k| z^k$, which will be denoted by $R(p)$. The first proof that $|\lambda| \leq R(p)$ is attributed to Cauchy [35]. This classical result is also proved in [165] as a corollary of Pellet's theorem and a recent proof can be found in [76, p.14]. Note that the fact that $u(z)$ has a unique positive real root, whenever $a_i \neq 0$ for at least one $i \in \{0, 1, \ldots, n-1\}$, follows from Descartes's rule of signs. Among all bounds on $|\lambda|$ that depend only on $|a_i|$, for $i = 0, 1, \ldots, n-1$, the sharpest one is precisely $R(p)$. This was stated in [165] and it is proved in Theorem 6.14 for completeness.

Theorem 6.14. [165, p.61] *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with $a_i \neq 0$ for at least one $i \in \{0, 1, \ldots, n-1\}$, and let $R(p)$ be the unique positive real root of $u(z) = z^n - \sum_{k=0}^{n-1} |a_k| z^k$. If $B(p)$ is an upper bound on the absolute values of the roots of $p(z)$ that is a function only of $|a_0|, |a_1|, \ldots, |a_{n-1}|$, then $R(p) \leq B(p)$.*

Proof. Since $B(p)$ depends only on $|a_i|$, for $i = 0, 1, \ldots, n-1$, we have that $B(p) = B(u)$ and, since $R(p)$ is a root of $u(z)$ and is positive, we have that $R(p) \leq B(u) = B(p)$. □

The optimality of $R(p)$ makes it very interesting for the theoretical purpose of testing the quality of other upper bounds for $|\lambda|$ that depend only on the absolute values of the coefficients of the polynomial. However, $R(p)$ has a limited practical interest since its computation requires to compute the root of a polynomial[2]. In the context of this section, the optimal bound $R(p)$ is used in Theorem 6.16, which establishes that for all Fiedler companion matrices of $p(z)$ the optimal upper bound that can be obtained by using the ∞-norm and diagonal similarities is, in all cases, precisely $R(p)$. However, this result is again mainly of theoretical interest, since there is not an easy way of choosing the optimal diagonal similarity.

The proof of Theorem 6.16 requires to use one lemma and Proposition 2.20. Lemma 6.15 merges Theorem 1, Corollary 1, and Corollary 2 in [145]. The concepts mentioned in the statement of Lemma 6.15 are contained in [87]. Also, note that all the inequalities containing vectors should be understood componentwise.

Lemma 6.15. *Let $A = (a_{ij}) \in \mathbb{C}^{n \times n}$ and let $\rho(|A|)$ be the spectral radius of $|A| = (|a_{ij}|)$. Then:*

(a)
$$\inf_{D \text{ diagonal}} \|D^{-1} A D\|_\infty = \rho(|A|).$$

(b) *There exists a vector $x = (x_i) > 0$ such that $|A|x - \rho(|A|)x \leq 0$ if and only if*
$$\min_{D \text{ diagonal}} \|D^{-1} A D\|_\infty = \rho(|A|).$$
In this case, the minimum is attained at $D' = \operatorname{diag}(x) := \operatorname{diag}(x_1, \ldots, x_n)$.

(c) *If A is irreducible, then* (b) *holds and the minimum is attained in the right positive eigenvector x of $|A|$ corresponding to $\rho(|A|)$, i.e., in the right Perron vector of $|A|$.*

Recall that Proposition 2.20 states that a Fiedler matrix $M_\sigma(p)$ is an irreducible matrix if and only if $p(0) \neq 0$. Now, we are in the position of proving the main result of this section.

Theorem 6.16. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with with $a_i \neq 0$ for at least one $i \in \{0, 1, \ldots, n-1\}$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let $M_\sigma(p)$ be the Fiedler matrix of $p(z)$ associated with σ, and let $R(p)$ be the unique positive real root of $u(z) = z^n - \sum_{k=0}^{n-1} |a_k| z^k$. Then*

[2] Although this root is a very special one, and fast methods for computing it can be easily devised.

(a) *$R(p)$ is the spectral radius of $|M_\sigma(p)|$.*

(b) *We have*

$$\inf_{D\,\text{diagonal}} \|D^{-1}M_\sigma(p)D\|_\infty = R(p).$$

(c) *Moreover, if $a_0 \neq 0$ and if we denote by $x_\sigma(p) \in \mathbb{R}^n$ the right Perron vector of $|M_\sigma(p)|$, then*

$$\min_{D\,\text{diagonal}} \|D^{-1}M_\sigma(p)D\|_\infty = R(p), \tag{6.16}$$

and the minimum is attained at $D' = \operatorname{diag}(x_\sigma(p))$.

Proof. By Theorem 2.19, we have that $u(z) = z^n - \sum_{k=0}^{n-1} |a_k| z^k$ is the characteristic polynomial of the nonnegative matrix $|M_\sigma(p)| = M_\sigma(u)$. The discussion at the beginning of this section implies that $R(p) \geq |\nu|$ for any other root ν of $u(z)$, i.e., for any eigenvalue of $M_\sigma(u)$. This proves part (a). Part (b) follows from Lemma 6.15-(a). Finally, part (c) follows from Lemma 6.15-(c) and Proposition 2.20. □

Theorem 6.16-(b) does not guarantee that the infimum is attained and does not explain how to find an optimal diagonal similarity if $a_0 = 0$. However, in the case of the first Frobenius companion matrix $C_1(p)$ this problem can be easily fixed. This is shown in Proposition 6.17.

Proposition 6.17. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with $a_i \neq 0$ for at least one $i \in \{0, 1, \ldots, n-1\}$, let $C_1(p)$ be the first Frobenius companion matrix of $p(z)$, and let $R(p)$ be the unique positive root of $u(z) = z^n - \sum |a_k| z^k$. If $D = \operatorname{diag}(R(p)^{n-1}, \ldots, R(p), 1)$, then*

$$\|D^{-1}C_1(p)D\|_\infty = R(p).$$

Proof. If $x^T = \begin{bmatrix} R(p)^{n-1} & \cdots & R(p) & 1 \end{bmatrix}$, then it may be checked that $|C_1(p)|x = R(p)x$. Since $x > 0$, Theorem 6.16-(a) and Lemma 6.15-(b) imply the result. □

It is natural to conjecture that a result similar to Proposition 6.17 also holds for any Fiedler matrix just by replacing $\begin{bmatrix} R(p)^{n-1} & \cdots & R(p) & 1 \end{bmatrix}^T$ by the corresponding right Perron vector. However, Example 6.18 shows that this is not true, since the right Perron vectors of the entrywise absolute values of Fiedler matrices different that the first Frobenius companion matrices may have zero entries when $a_0 = 0$ and, so, we cannot apply Lemma 6.15-(b) based on the Perron vectors.

Example 6.18. Consider the four Fiedler matrices associated with a polynomial $p(z) = z^3 + a_2 z + a_1 z + a_0$ with $a_i \neq 0$ for at least one $i \in \{0, 1, 2\}$, that is,

$$M_{\sigma_1}(p) = \begin{bmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad M_{\sigma_2}(p) = \begin{bmatrix} -a_2 & 1 & 0 \\ -a_1 & 0 & 1 \\ -a_0 & 0 & 0 \end{bmatrix},$$
$$M_{\sigma_3}(p) = \begin{bmatrix} -a_2 & -a_1 & 1 \\ 1 & 0 & 0 \\ 0 & -a_0 & 0 \end{bmatrix}, \quad \text{and} \quad M_{\sigma_4}(p) = \begin{bmatrix} -a_2 & 1 & 0 \\ -a_1 & 0 & -a_0 \\ 1 & 0 & 0 \end{bmatrix},$$

and let $R(p)$ be the unique positive root of $u(z) = z^3 - |a_2|z^2 - |a_1|z - |a_0|$. It may be checked that the eigenvectors of $|M_{\sigma_1}(p)|, |M_{\sigma_2}(p)|, |M_{\sigma_3}(p)|, |M_{\sigma_4}(p)|$ associated with $R(p)$ are, respectively,

$$x_{\sigma_1}(p) = \begin{bmatrix} R(p)^2 \\ R(p) \\ 1 \end{bmatrix}, \quad x_{\sigma_2}(p) = \begin{bmatrix} 1 \\ R(p) - |a_2| \\ R(p)^2 - |a_2|R(p) - |a_1| \end{bmatrix},$$

$$x_{\sigma_3}(p) = \begin{bmatrix} R(p) \\ 1 \\ R(p)^2 - |a_2|R(p) - |a_1| \end{bmatrix} \quad \text{and} \quad x_{\sigma_4}(p) = \begin{bmatrix} R(p) \\ R(p)^2 - |a_2|R(p) \\ 1 \end{bmatrix},$$

which have nonnegative entries as a consequence of $u(R(p)) = 0$. If we denote by D_1, D_2, D_3, D_4 the diagonal matrices $\mathrm{diag}(x_{\sigma_1}), \mathrm{diag}(x_{\sigma_2}), \mathrm{diag}(x_{\sigma_3}), \mathrm{diag}(x_{\sigma_4})$ respectively, then D_1 is the only one that is nonsingular for any values of a_2, a_1, and a_0. For example, consider the monic polynomial of degree 3 with $a_0 = a_1 = 0$ and $a_2 \neq 0$. Then

$$x_{\sigma_1}(p) = \begin{bmatrix} |a_2|^2 \\ |a_2| \\ 1 \end{bmatrix}, \quad x_{\sigma_2}(p) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad x_{\sigma_3}(p) = \begin{bmatrix} |a_2| \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad x_{\sigma_4}(p) = \begin{bmatrix} |a_2| \\ 0 \\ 1 \end{bmatrix}.$$

Explicit formulas for the eigenvectors of Fiedler matrices are available in the literature (see [45] or Theorem 8.2 for a new way to obtain them), and this allows us to add further conditions on the coefficients of the polynomial under which Proposition 6.17 can be extended to other Fiedler matrices when $a_0 = 0$. Since the general case is messy, we limit ourselves in Proposition 6.19 to the Fiedler matrix $F(p)$ in (6.9) that has played a very relevant role in this chapter.

Proposition 6.19. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial with $a_0 = 0$ and $a_1 \neq 0$, let $F(p)$ be the Fiedler matrix of $p(z)$ defined in* (6.9)*, and let $R(p)$ be the unique positive root of $u(z) = z^n - \sum_{k=0}^{n-1} |a_k| z^k$. If $D = \mathrm{diag}(R(p), R(p)u_1(R(p)), \ldots, R(p)u_{n-2}(R(p)), 1)$, then*

$$\|D^{-1}F(p)D\|_\infty = R(p),$$

where, for $k = 1, 2, \ldots, n-2$, $u_k(z)$ is the degree-k Horner shift of $u(z)$ (see Definition 2.6*).*

Proof. For brevity, we denote $R = R(p)$ in the proof. Let $x := \begin{bmatrix} R & Ru_1(R) & \cdots & Ru_{n-2}(R) & 1 \end{bmatrix}^T$. Then, it is easy to check that $|F(p)|x = Rx$, i.e., x is the Perron right vector of $|F(p)|$. Next, we prove that $x > 0$. To this purpose, observe that the Horner shifts satisfy $u_k(z) = zu_{k-1}(z) - |a_{n-k}|$, for $k = 1, 2, \ldots, n$, and that $u_n(z) = u(z)$. Since $R > 0$ is a root of $u(z)$, the equation $u(z) = zu_{n-1}(z) - |a_0|$, together with $a_0 = 0$, imply $u_{n-1}(R) = 0$. Also, since $a_1 \neq 0$, we have $0 = u_{n-1}(R) = Ru_{n-2}(R) - |a_1|$ which implies $u_{n-2}(R) = |a_1|/R > 0$. With this, the recurrence relation $Ru_{k-1}(R) = u_k(R) + |a_{n-k}|$ implies $u_j(R) > 0$, for $j = n-3, n-4, \ldots, 1$. Therefore, the Perron vector x is a positive vector and Theorem 6.16-(a) and Lemma 6.15-(b) imply the result. □

One point that should be remarked on Theorem 6.16-(c) is related to the fact mentioned above that, for a given eigenvalue of any Fiedler matrix, there exists a formula for the corresponding eigenvector (see [45] or Theorem 8.2). This formula depends, of course, on the eigenvalue and also on the Horner shifts of the polynomial, and is particularly simple in the cases of the Frobenius companion matrices. A potential use of these formulas is to obtain "approximately optimal" diagonal matrices to be used in $\|D^{-1}M_\sigma(p)D\|_\infty$. The idea would be to obtain first an upper bound on the absolute values of the roots of a polynomial by some of the approaches explained in this manuscript, to introduce this bound in the formula for the eigenvector of the corresponding Fiedler matrix $M_\sigma(u)$ for getting a vector y, and to take $D = \mathrm{diag}(y)$. This process can be iterated. This and other approaches for getting good bounds via diagonal similarities will be investigated in the near future.

Another interesting point to be commented is that a similar approach to the one explained in this section is possible for the inverses of Fiedler matrices. We do not present here all the details, but just the main ideas. Note that by Theorem 3.2, we have $|M_\sigma(p)^{-1}| = M_\sigma(l)^{-1}$, where $l(z) = z^n + \sum_{k=1}^{n-1} |a_k| z^k - |a_0|$, and, moreover, $l(z)$ has a unique positive real root [165], that we denote by $r(p)$. In addition, a nonsingular matrix is irreducible if and only if its inverse is irreducible. Therefore,

$r(p)^{-1}$ is the Perron eigenvalue of $|M_\sigma(p)^{-1}|$ and $\min_{D\,\mathrm{diag}} \|D^{-1}M_\sigma(p)^{-1}D\|_\infty = r(p)^{-1}$. Finally, note that the developments in this section, and the corresponding ones for inverses of Fiedler matrices, can be applied to the Fiedler matrices of $p^\sharp(z)$ and their inverses and, therefore, the diagonal similarities of all lower and upper bounds in (6.1) and (6.6) for the 1-norm are covered.

Chapter 7

Condition numbers for inversion of Fiedler matrices

As we said in Section 1.3, Frobenius companion matrices possess several properties that are undesirable numerically, and some of these properties are a consequence of having large condition number for inversion. For this reason we investigate in this chapter the condition numbers for inversion of Fiedler matrices, with the purpose of comparing them and to provide a simple criterion that allows us to determine in advance which Fiedler matrices have the smallest condition number.

The first point to be remarked is that there are no explicit expressions for the singular values of other Fiedler matrices than the Frobenius companion matrices (see Chapter 4), which prevents the use of the spectral norm in our developments. We have used instead the Frobenius norm which satisfies $\kappa_2(A) \leq \kappa_F(A) \leq \sqrt{n}\kappa_2(A)$ and $n \leq \kappa_F(A)$, in contrast with $1 \leq \kappa_2(A)$. These inequalities point out that studying ratios of condition numbers using the 2-norm or the Frobenius norm is essentially equivalent from a practical point of view.

7.1 Condition numbers for inversion of Fiedler matrices

We start by presenting in Theorem 7.1 an explicit expression for the condition number of any Fiedler matrix in the Frobenius norm, as an immediate consequence of Corollary 3.5. This expression will allow us to easily establish several relevant properties of these condition numbers. The quantity t_σ, that is, the number of initial consecutions or inversions of σ (see Part (a) in Definition 2.8) will play an important role in this chapter.

Theorem 7.1. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let M_σ be the Fiedler matrix of $p(z)$ associated with σ, and let t_σ be the number of initial consecutions or inversions of σ. Define*

$$N(p)^2 := (n-1) + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2.$$

Then,

$$\kappa_F^2(M_\sigma) = N(p)^2 \left((n-1) + \frac{1 + |a_1|^2 + \cdots + |a_{t_\sigma}|^2}{|a_0|^2} + |a_{t_\sigma+1}|^2 + \cdots + |a_{n-1}|^2 \right). \tag{7.1}$$

Corollary 7.2 gives crude lower and upper bounds on $\kappa_F(M_\sigma)$ that are independent on σ and show that, for any σ, $\kappa_F(M_\sigma)$ is large if and only if $|a_0|$ is small or $|a_i|$ is large for some $i = 0, 1, \ldots, n-1$ (or both).

Corollary 7.2. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ.*

(a) *If $|a_0| \leq 1$, then*

$$\frac{\sqrt{(n-1) + |a_1|^2 + \cdots + |a_{n-1}|^2}}{|a_0|} \leq \kappa_F(M_\sigma) \leq \frac{n + |a_1|^2 + \cdots + |a_{n-1}|^2}{|a_0|} .$$

(b) *If $|a_0| > 1$, then*

$$\sqrt{(n-1) + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2} \leq \kappa_F(M_\sigma) \leq (n-1) + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2 .$$

The proof of Corollary 7.2 is omitted since it follows trivially from Theorem 7.1. We would like to remark that it is natural that $\kappa_F(M_\sigma)$ is large if $|a_0|$ is small, because M_σ is singular when $a_0 = 0$. However, it might not be so clear why $\kappa_F(M_\sigma)$ is large, i.e., M_σ is close in relative distance to a singular matrix, if $|a_i|$ is large for some $i = 0, 1, \ldots, n-1$. The reason relies on Theorem 2.19-(d), because if some $|a_i| \gg 1$, then a tiny relative normwise perturbation can turn one of the entries equal to 1 in M_σ into 0 and can make the matrix singular. This property shows that "representing" a polynomial $p(z)$ via a Fiedler companion matrix is not convenient if some $|a_i| \gg 1$ because the "structural" entries equal to one are fragile under non-structured tiny perturbations.

Another direct consequence of Theorem 7.1 is Corollary 7.3, which gives a necessary and sufficient condition for two Fiedler matrices to have the same condition numbers for any monic polynomial $p(z)$.

Corollary 7.3. *Let $\mathbb{P}_n = \{z^n + \sum_{k=0}^{n-1} a_k z^k : a_0 \neq 0\}$ be the set of monic polynomials of degree $n \geq 2$ without roots equal to 0. Let $\sigma_1, \sigma_2 : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be two bijections and let t_{σ_1} and t_{σ_2} be, respectively, the numbers of initial consecutions or inversions of σ_1 and σ_2. Let $M_{\sigma_1}(p)$ and $M_{\sigma_2}(p)$ be, respectively, the Fiedler matrices of $p \in \mathbb{P}_n$ associated with σ_1 and σ_2. Then, $t_{\sigma_1} = t_{\sigma_2}$ if and only if $\kappa_F(M_{\sigma_1}(p)) = \kappa_F(M_{\sigma_2}(p))$ for all $p \in \mathbb{P}_n$.*

Proof. It is obvious that $t_{\sigma_1} = t_{\sigma_2}$ implies $\kappa_F(M_{\sigma_1}(p)) = \kappa_F(M_{\sigma_2}(p))$ for all $p \in \mathbb{P}_n$ by (7.1). To prove the converse, assume that $\kappa_F(M_{\sigma_1}(p)) = \kappa_F(M_{\sigma_2}(p))$ for all $p \in \mathbb{P}_n$ and proceed by contradiction, i.e., assume $t_{\sigma_1} \neq t_{\sigma_2}$. More precisely assume without loss of generality that $t_{\sigma_1} < t_{\sigma_2}$. Take $p(z)$ such that $a_0 = 2$, $a_{t_{\sigma_2}} = 1$, and $a_i = 0$ for $i \neq 0, t_{\sigma_2}$. Then $\kappa_F(M_{\sigma_1}(p)) = \kappa_F(M_{\sigma_2}(p))$ and (7.1) imply $1/4 + 1 = (1+1)/4$, which is a contradiction. □

Example 7.4. In this example all considered Fiedler matrices correspond to the same polynomial $p(z)$. The condition numbers in Frobenius norm of the Frobenius companion matrices C_1 and C_2 are equal. This is obvious because $C_2 = C_1^T$. It is however somewhat surprising that the condition numbers of the two pentadiagonal Fiedler matrices P_1 and P_2 introduced in (2.7) are, in general, different. This follows from Corollary 7.3 and the fact $t_{\sigma_1} = 1$ for P_1 and $t_{\sigma_2} = 2$ for P_2 (see Section 2.2). In fact, we will see in Theorem 7.10 that these condition numbers can be arbitrarily different for properly chosen polynomials.

7.2 Ordering Fiedler matrices according to condition numbers in the Frobenius norm

Given a monic polynomial $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ of degree $n \geq 2$ with $a_0 \neq 0$, and a number t such that $1 \leq t \leq n-1$, Corollary 7.3 establishes that all Fiedler matrices of $p(z)$ in the set

$$\mathcal{S}_t(p) := \{M_\sigma(p) : t_\sigma = t\}$$

have the same condition number $\kappa_F(M_\sigma(p))$. In the generic case $a_0 \neq -1$ (recall Corollary 2.18) the cardinality of $\mathcal{S}_t(p)$ is given by

$$|\mathcal{S}_t(p)| = \begin{cases} 2^{n-1-t}, & \text{if } t < n-1, \\ 2, & \text{if } t = n-1. \end{cases} \tag{7.2}$$

This can be seen as follows. If $t_\sigma = n-1$, then σ has $n-1$ consecutions and no inversions, or vice versa. This corresponds to the two classical Frobenius companion matrices. If $t_\sigma = t < n-1$, then σ has consecutions at $0, 1, \ldots, t-1$ and an inversion at t, or vice versa. For each of these two cases, we can select freely the consecutions/inversions at $t+1, \ldots, n-2$. This can be done in 2^{n-2-t} different ways, that according to `Algorithm 1` in Theorem 2.16 give each of them a different Fiedler matrix. The value of t in $\mathcal{S}_t(p)$ and expression (7.1) allow us to order all Fiedler matrices of $p(z)$ by increasing/decreasing condition numbers in Corollary 7.5. *Observe that there are only three possible different orders of this type, since the order via increasing/decreasing condition numbers is the same for all polynomials with $|p(0)| < 1$, it is also the same for all polynomials with $|p(0)| > 1$, and also the same for all polynomials with $|p(0)| = 1$.*

Corollary 7.5. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, and let t be a number such that $1 \leq t \leq n-1$. Let $\mathcal{S}_t(p) = \{M_\sigma(p) : t_\sigma = t\}$ be the set of Fiedler matrices of $p(z)$ associated with bijections σ whose number of initial consecutions or inversions is equal to t. Define*

$$\kappa(t) := \kappa_F(M_\sigma(p)), \quad \text{for } M_\sigma(p) \in \mathcal{S}_t(p), \tag{7.3}$$

which does not depend on the specific bijection σ as long as $t_\sigma = t$. Then the following results hold.

(a) *If $|a_0| < 1$, then $\kappa(1) \leq \kappa(2) \leq \cdots \leq \kappa(n-1)$.*

(b) *If $|a_0| = 1$, then $\kappa(1) = \kappa(2) = \cdots = \kappa(n-1)$.*

(c) *If $|a_0| > 1$, then $\kappa(1) \geq \kappa(2) \geq \cdots \geq \kappa(n-1)$.*

Proof. The result follows from (7.1), since this expression makes obvious that if $|a_0| < 1$, then $\kappa_F(M_\sigma(p))$ increases as the number t_σ of coefficients $|a_i|^2$ divided by $|a_0|^2$ increases. The other cases are proved in a similar way. □

Remark 7.6. From Corollary 7.5 we see that if $|a_0| < 1$, then the two Frobenius companion matrices have the largest condition number among all Fiedler matrices of $p(z)$, since the set $\mathcal{S}_{n-1}(p) = \{M_\sigma(p) : t_\sigma = n-1\}$ contains only the two Frobenius companion matrices. On the contrary, the Fiedler matrices in $\mathcal{S}_1(p) = \{M_\sigma(p) : t_\sigma = 1\}$ have the smallest condition number among all Fiedler matrices of $p(z)$ if $|a_0| < 1$. If n is large, then there are many Fiedler matrices with smallest condition number, since according to (7.2), $\mathcal{S}_1(p)$ has 2^{n-2} elements. In particular, $\mathcal{S}_1(p)$ contains the pentadiagonal Fiedler matrices P_1 and $P_3 = P_1^T$, in (2.7), but not the pentadiagonal matrices P_2 and $P_4 = P_2^T$, also in (2.7), which have a larger condition number if $|a_0| < 1$.

If $|a_0| > 1$, then similar remarks hold but with reverse order for the magnitudes of the condition numbers. In this case, the Frobenius companion matrices have the smallest condition number among all Fiedler matrices of $p(z)$.

The clear and simple ordering of Fiedler matrices according to condition numbers in the Frobenius norm presented in Corollary 7.5 does not hold for condition numbers in other matrix norms often used in the literature as, for instance, the $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ [87]. This is one of the reasons why we have chosen to use the Frobenius norm in this work. Of course, the equivalence of all these norms via constants smaller than or equal to n implies that the order in Corollary 7.5 between $\kappa(t)$ and $\kappa(t+1)$ can be broken in other norms only if $\kappa(t)$ and $\kappa(t+1)$ are not very different. We illustrate these points in Example 7.7, which also shows that an ordering based on the number of initial consecutions or inversions of σ is not possible for these other norms.

Example 7.7. In Figure 7.2.1 we consider the polynomials $p_1(z) = z^5 + 0.01z^4 + 0.01z^3 + 0.01z^2 + 0.01z + 0.01$ and $p_2(z) = z^5 + 10z^4 + z^3 + z^2 + 10z + 0.01$, both of degree 5. We have constructed with MATLAB the eight Fiedler matrices for each of these polynomials associated with bijections having an inversion at 0. The matrices associated with bijections having a consecution at 0 are the transposes of the previous ones and have not been considered for simplicity. Each of these Fiedler matrices has been labeled with an index from 1 to 8, according to the following table.

index	CISS(σ)	t_σ
1	$(0, 4)$	4
2	$(0, 3, 1, 0)$	3
3	$(0, 2, 1, 1)$	2
4	$(0, 2, 2, 0)$	2

index	CISS(σ)	t_σ
5	$(0, 1, 1, 2)$	1
6	$(0, 1, 1, 1, 1, 0)$	1
7	$(0, 1, 2, 1)$	1
8	$(0, 1, 3, 0)$	1

These indices are represented in the horizontal axes of the plots in Figure 7.2.1. For these 8 Fiedler matrices of each polynomial $p_1(z)$ and $p_2(z)$, we have computed their condition numbers in the 2-norm (that is, the ratio between the largest and smallest singular values) and we have ordered the matrices by decreasing magnitudes of these condition numbers, i.e., the matrix with the largest condition number is in the first position. The positions of the Fiedler matrices with respect this ordering are represented in the vertical axes of the plots in Figure 7.2.1 by using the symbol "•". In addition, the positions of the same Fiedler matrices with respect the ordering corresponding to decreasing Frobenius condition numbers are represented in the vertical axes by using the symbol "+". We observe that the ordering with respect the 2-norm condition number differs completely from $p_1(z)$ to $p_2(z)$, and in both cases is very different from the one corresponding to Frobenius condition numbers. Other interesting point to be remarked is that, both for $p_1(z)$ and $p_2(z)$, the condition numbers in the 2-norm of the eight considered Fiedler matrices are all different to each other, and so the same value of t_σ does not imply the same condition number in the 2-norm, by contrast with the behaviour in the Frobenius norm. This is not seen in Figure 7.2.1, but it may be easily checked by the reader with MATLAB. Finally, we mention that the condition numbers in the Frobenius norm for the Fiedler matrices of $p_1(z)$ range from 200.063 to 200.093, while the ones corresponding to $p_2(z)$ range from $1.443 \cdot 10^4$ to $2.045 \cdot 10^4$.

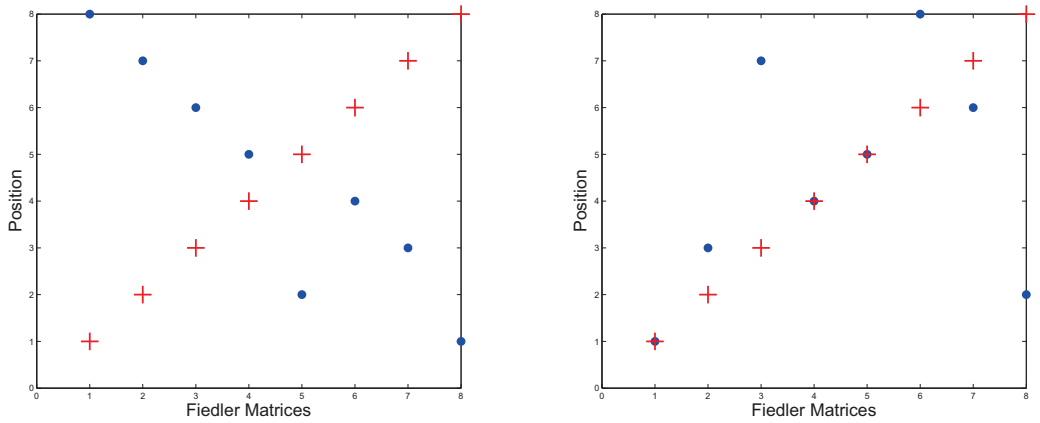
In Figure 7.2.2, we repeat the same experiment for the 1-norm instead of the 2-norm and for the polynomials $p_3(z) = z^5 + 10z^4 + 100z^3 + 10z^2 + 100z + 0.01$ and $p_4(z) = z^5 + 100z^4 + 10z^3 + 100z^2 + 10z + 0.01$. The results obtained are similar to those in the 2-norm. In this case, the condition numbers in the Frobenius norm for the Fiedler matrices of $p_3(z)$ range from $1.421 \cdot 10^6$ to $2.020 \cdot 10^6$, while the corresponding to $p_4(z)$ range from $2.020 \cdot 10^6$ to $0.144 \cdot 10^6$.

We do not show experiments in the ∞-norm, because the ∞-norm condition number of a matrix is the 1-norm condition number of its transpose, and the transpose of any Fiedler matrix is another Fiedler matrix with the same number of initial consecutions or inversions.

7.3 The ratio of the condition numbers of two Fiedler matrices

The important fact in applications is not whether one matrix is better conditioned than another. The really important fact is to know whether the condition number of one matrix is much smaller than the condition number of another or not. Therefore, we study in this section the ratio between the condition numbers in the Frobenius norm of any pair of Fiedler matrices of a fixed polynomial $p(z)$ that have different numbers of initial consecutions or inversions.

Lemma 7.8 states a simple technical result that will be used in the rest of this section.



(a) $p_1(z) = z^5 + 0.01z^4 + 0.01z^3 + 0.01z^2 + 0.01z + 0.01$ (b) $p_2(z) = z^5 + 10z^4 + z^3 + z^2 + 10z + 0.01$

Figure 7.2.1: Ordering Fiedler matrices of a fixed polynomial according to decreasing condition numbers in the 2-norm (•) and in the Frobenius norm (+).

Lemma 7.8. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $\sigma, \mu : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be two bijections, let M_σ and M_μ be the Fiedler matrices of $p(z)$ associated with σ and μ, and let t_σ and t_μ be the numbers of initial consecutions or inversions of σ and μ. Assume that $t_\sigma < t_\mu$ and define*

$$g_{\sigma,\mu} := (n-1) + \frac{1 + |a_1|^2 + \cdots + |a_{t_\sigma}|^2}{|a_0|^2} + |a_{t_\mu+1}|^2 + \cdots + |a_{n-1}|^2, \tag{7.4}$$

where if $t_\mu = n-1$, then $|a_{t_\mu+1}|^2 + \cdots + |a_{n-1}|^2$ is not present. Then

$$\left(\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\right)^2 = \frac{g_{\sigma,\mu} + \dfrac{|a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{|a_0|^2}}{g_{\sigma,\mu} + |a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}$$

and

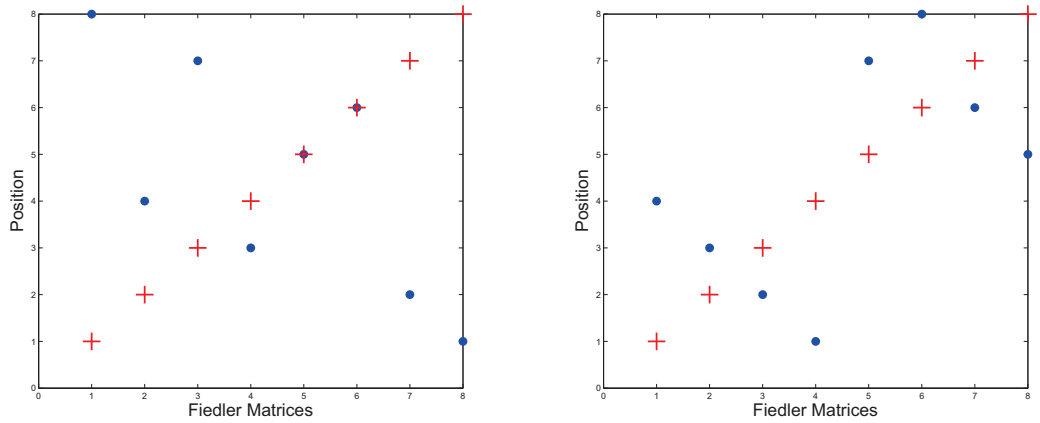
$$\left(\frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)}\right)^2 = \frac{g_{\sigma,\mu} + |a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{g_{\sigma,\mu} + \dfrac{|a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{|a_0|^2}}.$$

Proof. The result is another corollary of (7.1). Simply note that

$$\kappa_F^2(M_\mu) = N(p)^2 \left(g_{\sigma,\mu} + \frac{|a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{|a_0|^2}\right)$$

and $\kappa_F^2(M_\sigma) = N(p)^2 \left(g_{\sigma,\mu} + |a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2\right)$. □

The ratios between condition numbers in Lemma 7.8 are functions of the coefficients of the polynomial $p(z)$ defined through somewhat involved formulas. Theorem 7.9 provides simple upper bounds for these ratios, which show that distinct Fiedler matrices of the same polynomial $p(z)$ may have very different condition numbers only if some of the coefficients $a_2, a_3, \ldots, a_{n-1}$ of the polynomial is very large, and a_0 is very small or very large.



(a) $p_3(z) = z^5 + 10z^4 + 100z^3 + 10z^2 + 100z + 0.01$ (b) $p_4(z) = z^5 + 100z^4 + 10z^3 + 100z^2 + 10z + 0.01$

Figure 7.2.2: Ordering Fiedler matrices of a fixed polynomial according to decreasing condition numbers in the 1-norm (•) and in the Frobenius norm (+).

Theorem 7.9. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $\sigma, \mu : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be two bijections, let M_σ and M_μ be the Fiedler matrices of $p(z)$ associated with σ and μ, and let t_σ and t_μ be the numbers of initial consecutions or inversions of σ and μ. Assume that $t_\sigma < t_\mu$ and define*

$$S_{\sigma,\mu} := \sum_{i=t_\sigma+1}^{t_\mu} |a_i|^2 \quad \text{and} \quad A = \max_{2 \leq i \leq n-1} |a_i|. \tag{7.5}$$

Then, the following statements hold.

(a) *If $|a_0| < 1$, then*

$$1 \leq \frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)} \leq \min\left\{\sqrt{1 + S_{\sigma,\mu}}\,,\, \frac{1}{|a_0|}\right\} \leq \min\left\{\sqrt{1 + (n-2)A^2}\,,\, \frac{1}{|a_0|}\right\}. \tag{7.6}$$

(b) *If $|a_0| > 1$, then*

$$1 \leq \frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)} \leq \min\left\{\sqrt{1 + \frac{S_{\sigma,\mu}}{n-1}}\,,\, |a_0|\right\} \leq \min\left\{\sqrt{1 + \frac{n-2}{n-1}A^2}\,,\, |a_0|\right\}. \tag{7.7}$$

Observe that the rightmost upper bounds in parts (a) *and* (b) *are both independent of σ and μ.*

Proof. Part (a). From Corollary 7.5 and Lemma 7.8, we have

$$1 \leq \left(\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\right)^2 \leq \frac{g_{\sigma,\mu} + \dfrac{|a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{|a_0|^2}}{g_{\sigma,\mu}} = 1 + \frac{|a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{g_{\sigma,\mu}\,|a_0|^2}$$
$$\leq 1 + |a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2 = 1 + S_{\sigma,\mu}\,,$$

where in the last inequality we have used that $1 < g_{\sigma,\mu}\,|a_0|^2$. To get the rightmost bound in Part (a), recall that $1 \leq t_\sigma\,,\, t_\mu \leq (n-1)$. So $S_{\sigma,\mu} \leq (n-2)A^2$. Next, we bound the ratio of condition

numbers by $1/|a_0|$. To this purpose define $y := S_{\sigma,\mu}/g_{\sigma,\mu} \geq 0$ and $\alpha := 1/|a_0|^2 > 1$. Therefore Lemma 7.8 implies

$$\left(\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\right)^2 = \frac{1+\alpha\, y}{1+y}. \tag{7.8}$$

Observe that the function $g(y) = (1+\alpha\, y)/(1+y)$ satisfies: (i) $g(0) = 1$; (ii) $\lim_{y\to\infty} g(y) = \alpha$; and (iii) $g'(y) = (\alpha-1)/(1+y)^2 > 0$. Therefore, $1 \leq g(y) < \alpha$, if $y \geq 0$, and (7.8) implies that $\kappa_F(M_\mu)/\kappa_F(M_\sigma) < 1/|a_0|$.

Part (b). From Corollary 7.5 and Lemma 7.8, we have

$$1 \leq \left(\frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)}\right)^2 \leq \frac{g_{\sigma,\mu} + |a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{g_{\sigma,\mu}} = 1 + \frac{|a_{t_\sigma+1}|^2 + \cdots + |a_{t_\mu}|^2}{g_{\sigma,\mu}} \leq 1 + \frac{S_{\sigma,\mu}}{n-1},$$

where in the last inequality we have used that $n-1 < g_{\sigma,\mu}$. To get the rightmost bound in Part (b), we use again that $S_{\sigma,\mu} \leq (n-2)A^2$. Next, we bound the ratio of condition numbers by $|a_0|$. To this purpose define $y := S_{\sigma,\mu}/g_{\sigma,\mu} \geq 0$ and $\alpha := 1/|a_0|^2 < 1$. Therefore Lemma 7.8 implies

$$\left(\frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)}\right)^2 = \frac{1+y}{1+\alpha\, y}. \tag{7.9}$$

Observe that the function $h(y) = (1+y)/(1+\alpha\, y)$ satisfies: (i) $h(0) = 1$; (ii) $\lim_{y\to\infty} h(y) = 1/\alpha$; and (iii) $h'(y) = (1-\alpha)/(1+\alpha y)^2 > 0$. Therefore, $1 \leq h(y) < 1/\alpha$, if $y \geq 0$, and (7.9) implies that $\kappa_F(M_\sigma)/\kappa_F(M_\mu) < |a_0|$. □

It is obvious that there exist polynomials $p(z)$ for which the upper bound $\min\left\{\sqrt{1+S_{\sigma,\mu}},\, 1/|a_0|\right\}$ in (7.6) (resp. $\min\left\{\sqrt{1+S_{\sigma,\mu}/(n-1)},\, |a_0|\right\}$ in (7.7)) can be as large as desired, but this does not mean necessarily that the ratio $\kappa_F(M_\mu)/\kappa_F(M_\sigma)$ (resp. $\kappa_F(M_\sigma)/\kappa_F(M_\mu)$) for these polynomials is large. In fact, note that $\min\left\{\sqrt{1+S_{\sigma,\mu}},\, 1/|a_0|\right\}$ (resp. $\min\left\{\sqrt{1+S_{\sigma,\mu}/(n-1)},\, |a_0|\right\}$) does not depend of the coefficients of the polynomial that define the magnitude $g_{\sigma,\mu}$ in (7.4), with the exception of a_0. Therefore, according to Lemma 7.8, the upper bounds in (7.6) or (7.7) cannot determine the actual values of the ratios $\kappa_F(M_\mu)/\kappa_F(M_\sigma)$ or $\kappa_F(M_\sigma)/\kappa_F(M_\mu)$. Theorem 7.10 shows that if we fix a priori an arbitrary value of the upper bound in (7.6) or in (7.7), then there exist polynomials for which this upper bound is almost attained and another polynomials for which the ratios of the condition numbers of Fiedler matrices are arbitrarily close to 1. Note that, in particular, Theorem 7.10 shows that there are polynomials for which the ratios of the Frobenius condition numbers of two distinct Fiedler matrices can be arbitrarily large or arbitrarily small.

Theorem 7.10. *Let $\sigma, \mu : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be two bijections and let t_σ and t_μ be the numbers of initial consecutions or inversions of σ and μ. Assume that $t_\sigma < t_\mu$. For any monic polynomial $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ of degree $n \geq 2$, let $g_{\sigma,\mu}(p)$ be the expression in (7.4), let $S_{\sigma,\mu}(p)$ be the first expression in (7.5), and let $M_\sigma(p)$ and $M_\mu(p)$ be the Fiedler matrices of $p(z)$ associated with σ and μ. Let $\boldsymbol{b} > 1$ be a given real number and define the sets of polynomials*

$$\mathcal{L}_{\boldsymbol{b}} := \left\{ p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k \,:\, \min\left\{\sqrt{1+S_{\sigma,\mu}(p)},\, \frac{1}{|a_0|}\right\} = \boldsymbol{b},\ 0 \neq |a_0| < 1 \right\},$$

$$\mathcal{M}_{\boldsymbol{b}} := \left\{ p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k \,:\, \min\left\{\sqrt{1+\frac{S_{\sigma,\mu}(p)}{n-1}},\, |a_0|\right\} = \boldsymbol{b},\ 1 < |a_0| \right\}.$$

Then the following statements hold.

(a) *For all* $\epsilon > 0$*, there exists* $p(z) \in \mathcal{L}_{\boldsymbol{b}}$ *such that*

$$1 \leq \left(\frac{\kappa_F(M_\mu(p))}{\kappa_F(M_\sigma(p))}\right)^2 \leq 1+\epsilon.$$

In particular, this happens for any $p(z) \in \mathcal{L}_{\boldsymbol{b}}$ *whose coefficient* a_1 *satisfies* $S_{\sigma,\mu}(p)/\epsilon \leq |a_1|^2$*.*

(b) *For all* $\epsilon > 0$*, there exists* $p(z) \in \mathcal{L}_{\boldsymbol{b}}$ *such that*

$$\frac{\boldsymbol{b}^2}{1+\epsilon} \leq \left(\frac{\kappa_F(M_\mu(p))}{\kappa_F(M_\sigma(p))}\right)^2 \leq \boldsymbol{b}^2.$$

In particular, this happens for any $p(z) \in \mathcal{L}_{\boldsymbol{b}}$ *such that* $S_{\sigma,\mu}(p)$ *satisfies* $\max\{1/|a_0|^2, g_{\sigma,\mu}(p)/\epsilon\} \leq S_{\sigma,\mu}(p)$*. Note that in this case* $1/|a_0| = \boldsymbol{b}$*.*

(c) *For all* $\epsilon > 0$*, there exists* $p(z) \in \mathcal{M}_{\boldsymbol{b}}$ *such that*

$$1 \leq \left(\frac{\kappa_F(M_\sigma(p))}{\kappa_F(M_\mu(p))}\right)^2 \leq 1+\epsilon.$$

In particular, this happens for any $p(z) \in \mathcal{M}_{\boldsymbol{b}}$ *whose coefficient* a_1 *satisfies* $(|a_0|^2\, S_{\sigma,\mu}(p)/\epsilon) \leq |a_1|^2$*.*

(d) *For all* $\epsilon > 0$*, there exists* $p(z) \in \mathcal{M}_{\boldsymbol{b}}$ *such that*

$$\frac{\boldsymbol{b}^2}{1+\epsilon} \leq \left(\frac{\kappa_F(M_\sigma(p))}{\kappa_F(M_\mu(p))}\right)^2 \leq \boldsymbol{b}^2.$$

In particular, this happens for any $p(z) \in \mathcal{M}_{\boldsymbol{b}}$ *such that* $|a_0| = \boldsymbol{b}$ *and* $S_{\sigma,\mu}(p)$ *satisfies* $(|a_0|^2\, g_{\sigma,\mu}(p)/\epsilon) \leq S_{\sigma,\mu}(p)$*.*

Proof. Part (a). Let $p(z) \in \mathcal{L}_{\boldsymbol{b}}$ be such that its coefficient a_1 satisfies $S_{\sigma,\mu}(p)/\epsilon \leq |a_1|^2$. In the following developments all magnitudes refer to $p(z)$, but the dependence on $p(z)$ is dropped for simplicity. From Corollary 7.5 and Lemma 7.8, we get

$$1 \leq \left(\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\right)^2 = \frac{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}}{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}} \leq 1+\frac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2} \leq 1+\frac{S_{\sigma,\mu}}{|a_1|^2} \leq 1+\epsilon\,.$$

Part (b). Let $p(z) \in \mathcal{L}_{\boldsymbol{b}}$ be such that $S_{\sigma,\mu}(p)$ satisfies $\max\{1/|a_0|^2, g_{\sigma,\mu}(p)/\epsilon\} \leq S_{\sigma,\mu}(p)$. In the following developments all magnitudes refer to $p(z)$, but the dependence on $p(z)$ is dropped for simplicity. From Lemma 7.8 and Theorem 7.9, we get

$$\boldsymbol{b}^2 \geq \left(\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\right)^2 = \frac{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}}{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}} \geq \frac{\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}}{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}} \geq \frac{\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}}{\epsilon\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}} = \frac{\dfrac{1}{|a_0|^2}}{1+\epsilon} = \frac{\boldsymbol{b}^2}{1+\epsilon}\,.$$

Part (c). Let $p(z) \in \mathcal{M}_{\boldsymbol{b}}$ be such that its coefficient a_1 satisfies $(|a_0|^2\, S_{\sigma,\mu}(p)/\epsilon) \leq |a_1|^2$. In the following developments all magnitudes refer to $p(z)$, but the dependence on $p(z)$ is dropped for simplicity. From Corollary 7.5 and Lemma 7.8, we get

$$1 \leq \left(\frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)}\right)^2 = \frac{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}}{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}} \leq 1+\frac{S_{\sigma,\mu}}{g_{\sigma,\mu}} \leq 1+\frac{S_{\sigma,\mu}}{\dfrac{|a_1|^2}{|a_0|^2}} \leq 1+\epsilon\,.$$

Part (d). Let $p(z) \in \mathcal{M}_{\boldsymbol{b}}$ be such that $|a_0| = \boldsymbol{b}$ and $S_{\sigma,\mu}(p)$ satisfies $(|a_0|^2\, g_{\sigma,\mu}(p)/\epsilon) \leq S_{\sigma,\mu}(p)$. In the following developments all magnitudes refer to $p(z)$, but the dependence on $p(z)$ is dropped for simplicity. From Lemma 7.8 and Theorem 7.9, we get

$$\boldsymbol{b}^2 \geq \left(\frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)}\right)^2 = \frac{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}}{1+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}} \geq \frac{\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}}}{\epsilon\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}+\dfrac{S_{\sigma,\mu}}{g_{\sigma,\mu}\,|a_0|^2}} = \frac{|a_0|^2}{1+\epsilon} = \frac{\boldsymbol{b}^2}{1+\epsilon}\,.$$

□

Remark 7.11. We have shown in Theorem 7.10 how to easily construct families of polynomials where the upper bounds in Theorem 7.9 for the ratios of condition numbers of different Fiedler matrices of the same polynomial are essentially attained, and other families where they are far from being attained. The reader should keep in mind that there are other families of polynomials satisfying the same properties.

Corollary 7.5 and Theorem 7.10 suggest that for polynomials with $|p(0)| < 1$ one should avoid the use of the classical Frobenius companion matrices and to use, instead, Fiedler matrices with a number of initial consecutions or inversions equal to one, as for instance P_1 in (2.7) or F in (2.8). This would lead to use matrices with the smallest possible condition number that, in addition, for certain polynomials may be arbitrarily smaller than the condition numbers of other Fiedler matrices. For polynomials with $|p(0)| > 1$ the situation is the opposite, and Frobenius companion matrices are the best choice from the point of view of condition numbers for inversion. However, Theorem 7.12 tells us that, given a monic polynomial $p(z)$, if there are two distinct Fiedler matrices with very different condition numbers, then both matrices are very ill-conditioned. Therefore, different Fiedler matrices may have very different condition numbers but only in cases where these matrices are nearly singular. The reciprocal is not true, because there may be two different Fiedler matrices nearly singular but having exactly the same condition number, as it is shown in Corollaries 7.2 and 7.5-(b).

Theorem 7.12. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$ with $a_0 \neq 0$, let $\sigma, \mu : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be two bijections, let M_σ and M_μ be the Fiedler matrices of $p(z)$ associated with σ and μ, and let t_σ and t_μ be the numbers of initial consecutions or inversions of σ and μ. Assume that $t_\sigma < t_\mu$. Then the following results hold.*

(a) *If $|a_0| < 1$, then*

$$1 \leq \left(\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\right)^2 \leq \kappa_F(M_\sigma) \leq \kappa_F(M_\mu)\,.$$

(b) *If $|a_0| > 1$, then*

$$1 \leq \frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)} \leq \kappa_F(M_\mu) \leq \kappa_F(M_\sigma)\,.$$

Proof. Part (a). From Theorem 7.10 and with the notation used there, we get

$$1 \leq \left(\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\right)^2 = \frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)}\,\frac{\kappa_F(M_\mu)}{\kappa_F(M_\sigma)} \leq \sqrt{1+S_{\sigma,\mu}}\,\frac{1}{|a_0|} \leq \|M_\sigma\|_F\,\|M_\sigma^{-1}\|_F\,,$$

where the last inequality follows from Corollary 3.5.

Part (b). From Theorem 7.10 and with the notation used there, we get

$$1 \leq \frac{\kappa_F(M_\sigma)}{\kappa_F(M_\mu)} \leq |a_0| \leq \|M_\sigma\|_F\,\|M_\sigma^{-1}\|_F\,,$$

where the last inequality follows again from Corollary 3.5. □

Remark 7.13. The difference between the statements of parts (a) and (b) in Theorem 7.12 is striking, but the next example shows that $(\kappa_F(M_\sigma)/\kappa_F(M_\mu))^2$ cannot be used in part (b). Consider the monic polynomial

$$p(z) = 10^4 + 2\,z + 2\,z^2 + 2 \cdot 10^5\,z^3 + 2 \cdot 10^5\,z^4 + 2\,z^5 + 2\,z^6 + 3\,z^7 + z^8$$

and the bijections σ and μ with consecution-inversion structure sequences $\mathrm{CISS}(\sigma) = (2, 1, 1, 2, 1, 0)$ and $\mathrm{CISS}(\mu) = (4, 2, 1, 0)$. In this case, $\kappa_F(M_\mu) = 8.124 \cdot 10^6$, $\kappa_F(M_\sigma) = 8.005 \cdot 10^{10}$, and $(\kappa_F(M_\sigma)/\kappa_F(M_\mu))^2 = 9.709 \cdot 10^7$. However, we see in this example that $(\kappa_F(M_\sigma)/\kappa_F(M_\mu)) << \kappa_F(M_\mu)$. We have observed the same behavior in all the examples that we have tested with large values of $\kappa_F(M_\sigma)/\kappa_F(M_\mu)$. Therefore, we think that the result in part (b) of Theorem 7.12 can be considerably improved.

Chapter 8

Pseudospectra and eigenvalue condition numbers of Fiedler matrices

In the numerical solution of a given problem, the forward error of a computed quantity (approximate solution) is the difference (in relative or absolute terms) between this quantity an the exact solution of the problem. To measure the accuracy of the numerical method one needs to get sharp bounds for the forward error. A basic inequality that relates the forward error with the condition number and the backward error is [79, p. 97]:

$$\text{forward error} \lesssim \text{condition number} \times \text{backward error}.$$

Then, in order to bound the forward error of the polynomial root finding problem solved as an eigenvalue problem we need to compare both the conditioning and the backward error of both problems. The analysis of the backward error of the polynomial root finding problem solved as an eigenvalue problem is carried out in Chapter 9.

In this chapter, we compare the condition number of a given root of a monic polynomial $p(z)$ with the condition number of this root as an eigenvalue of any Fiedler matrix, and the pseudozero sets of $p(z)$ with the pseudospectra of the associated Fiedler matrices (recall from Section 1.2.2.2 that pseudozero sets and pseudospectra are tools that give insight into the sensitivity of the roots of polynomials and the eigenvalues of matrices, respectively, to perturbations). In particular, in Section 8.1 we present expressions for the condition numbers of the roots of $p(z)$. In Section 8.2 we give explicit formulas for the right and left eigenvectors of Fiedler matrices since they will be needed to compute the eigenvalue condition numbers, and, then, in Section 8.3 we present expressions for the eigenvalue condition numbers of Fiedler matrices associated with $p(z)$. In Section 8.4 we compare the eigenvalue condition numbers of Fiedler matrices with the condition numbers of the roots of $p(z)$, and, also, we compare the eigenvalue condition numbers of the Frobenius companion matrices with the eigenvalue condition numbers of Fiedler matrices other than Frobenius ones. Section 8.5 is devoted to the study of pseudospectra of Fiedler matrices, and to compare them with the pseusozero sets of $p(z)$. Finally, in Section 8.6 we present numerical experiments to illustrate our theoretical results, and to study the effect of balancing Fiedler matrices from the point of view of eigenvalue condition numbers and pseudospectra.

As in Section 1.2.2, in order to better express our results, we need to distinguish between norms on the vector space of polynomials of degree less than or equal to n and norms on the vector space of coefficients of monic polynomials (excluding the leading coefficient $a_n = 1$) of degree equal to n. In particular, for a polynomial $p(z) = \sum_{k=0}^{n} a_k z^k$ non necessarily monic, $\|p\|_2$ is the norm on

the vector space of polynomials of degree less than or equal to n defined as

$$\|p\|_2 = \sqrt{\sum_{k=0}^{n} |a_k|^2}.$$

In addition, for a monic polynomial $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$, we define $|||p|||_2$ as

$$|||p|||_2 = \sqrt{\sum_{k=0}^{n-1} |a_k|^2}.$$

Notice that $|||p|||_2$ is not a norm on the vector space of polynomials of degree less than or equal to n. Also notice that, since we deal in this chapter with monic polynomials, we always have $\|p\|_2 \geq 1$.

8.1 Condition numbers of roots of monic polynomials

Recall from Section 1.2.2.1 that the condition number $\kappa(\lambda, p)$ and the coefficientwise condition number $\mathrm{cond}(\lambda, p)$ of a nonzero simple root λ of a monic polynomial $p(z)$ are, respectively,

$$\kappa(\lambda, p) := \lim_{\epsilon \to 0} \sup \left\{ \frac{|\widetilde{\lambda} - \lambda|}{\epsilon |\lambda|} : \widetilde{\lambda} \text{ is a root of } \widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k \text{ with } |||\widetilde{p} - p|||_2 \leq \epsilon |||p|||_2 \right\}, \tag{8.1}$$

and

$$\mathrm{cond}(\lambda, p) := \lim_{\epsilon \to 0} \sup \left\{ \frac{|\widetilde{\lambda} - \lambda|}{\epsilon |\lambda|} : \widetilde{\lambda} \text{ is a root of } \widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k \text{ with } \sqrt{\sum_{k=0}^{n-1} \left| \frac{\widetilde{a}_k - a_k}{a_k} \right|^2} \leq \epsilon \right\}. \tag{8.2}$$

In Theorem 8.1 we derive computable formulas for the condition numbers $\kappa(\lambda, p)$ and $\mathrm{cond}(\lambda, p)$.

Theorem 8.1. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let λ be a simple nonzero root of $p(z)$, let $\Lambda(z)$ and $\widehat{\Lambda}(z)$ be the vectors*

$$\Lambda(z) = \begin{bmatrix} z^{n-1} & \cdots & z & 1 \end{bmatrix}^T \quad \text{and} \quad \widehat{\Lambda}(z) = \begin{bmatrix} z^{n-1} a_{n-1} & \cdots & z a_1 & a_0 \end{bmatrix}^T, \tag{8.3}$$

and let $\kappa(\lambda, p)$ and $\mathrm{cond}(\lambda, p)$ be the condition numbers defined in (8.1) *and* (8.2)*, respectively. Then,*

$$\kappa(\lambda, p) = \frac{|||p|||_2 \|\Lambda(\lambda)\|_2}{|\lambda| \cdot |p'(\lambda)|} \quad \text{and} \quad \mathrm{cond}(\lambda, p) = \frac{\|\widehat{\Lambda}(\lambda)\|_2}{|\lambda| \cdot |p'(\lambda)|}. \tag{8.4}$$

Proof. We only prove that $\kappa(\lambda, p) = |||p|||_2 \|\Lambda(\lambda)\|_2 / (|\lambda| \cdot |p'(\lambda)|)$, since a proof for $\mathrm{cond}(\lambda, p) = \|\widehat{\Lambda}(\lambda)\|_2 / (|\lambda| \cdot |p'(\lambda)|)$ can be obtained from the following proof with some minor modifications.

We first show that $|||p|||_2 \|\Lambda(\lambda)\|_2 / (|\lambda| \cdot |p'(\lambda)|)$ is an upper bound for $\kappa(\lambda, p)$. For this, consider a polynomial $\widetilde{p}(z)$ such that $|||\widetilde{p} - p|||_2 \leq \epsilon |||p|||_2$, and write

$$\begin{aligned} 0 = p(\lambda) &= p(\lambda) - p(\widetilde{\lambda}) + p(\widetilde{\lambda}) - \widetilde{p}(\widetilde{\lambda}) + \widetilde{p}(\widetilde{\lambda}) \\ &= p(\lambda) - p(\widetilde{\lambda}) + p(\widetilde{\lambda}) - \widetilde{p}(\widetilde{\lambda}) \\ &= (\lambda - \widetilde{\lambda}) p'(\lambda) + p(\lambda) - \widetilde{p}(\lambda) + (\widetilde{\lambda} - \lambda)(p'(\lambda) - \widetilde{p}'(\lambda)) + O(|\lambda - \widetilde{\lambda}|^2), \end{aligned}$$

where we have used that $\widetilde{\lambda}$ is a root of $\widetilde{p}(z)$. Then, since $p'(\lambda) \neq 0$,

$$\frac{|\lambda - \widetilde{\lambda}|}{\epsilon |\lambda|} = \frac{|p(\lambda) - \widetilde{p}(\lambda) + (\widetilde{\lambda} - \lambda)(p'(\lambda) - \widetilde{p}'(\lambda)) + O(|\lambda - \widetilde{\lambda}|^2)|}{\epsilon |\lambda| \cdot |p'(\lambda)|}$$

$$\leq \frac{|p(\lambda)-\widetilde{p}(\lambda)|}{\epsilon|\lambda|\cdot|p'(\lambda)|}+\frac{|\widetilde{\lambda}-\lambda|\cdot|p'(\lambda)-\widetilde{p}'(\lambda)|}{\epsilon|\lambda|\cdot|p'(\lambda)|}+\frac{O(|\lambda-\widetilde{\lambda}|^2)}{\epsilon|\lambda|\cdot|p'(\lambda)|}$$
$$\leq \frac{|||\widetilde{p}-p|||_2\|\Lambda(\lambda)\|_2}{\epsilon|\lambda|\cdot|p'(\lambda)|}+\frac{(n-1)|\lambda-\widetilde{\lambda}|\cdot|||\widetilde{p}-p|||_2\|\Lambda(\lambda)\|_2}{\epsilon|\lambda|\cdot|p'(\lambda)|}+\frac{O(|\lambda-\widetilde{\lambda}|^2)}{\epsilon|\lambda|\cdot|p'(\lambda)|}$$
$$\leq \frac{|||p|||_2\|\Lambda(\lambda)\|_2}{|\lambda|\cdot|p'(\lambda)|}+O(|\lambda-\widetilde{\lambda}|)\xrightarrow[\epsilon\longrightarrow 0]{}\frac{|||p|||_2\|\Lambda(\lambda)\|_2}{|\lambda|\cdot|p'(\lambda)|},$$

where we have used that $|p(\lambda)-\widetilde{p}(\lambda)|\leq\sum_{k=0}^{n-1}|a_k-\widetilde{a}_k|\cdot|\lambda^k|\leq|||p-\widetilde{p}|||_2\|\Lambda(\lambda)\|_2$ (by Cauchy-Schwarz inequality), and that $|||p'-\widetilde{p}'|||\leq(n-1)|||p-\widetilde{p}|||$.

Now, we prove that the supremum is attained at $|||p|||_2\|\Lambda(\lambda)\|_2/(|\lambda|\cdot|p'(\lambda)|)$, that is, given ϵ small enough, there are $\widetilde{p}(z)$ and $\widetilde{\lambda}$ satisfying:

$$\text{(i) } \widetilde{p}(\widetilde{\lambda})=0,\quad \text{(ii) } |||\widetilde{p}-p|||_2\leq\epsilon|||p|||_2,\quad \text{and}\quad \text{(iii) } \frac{|\lambda-\widetilde{\lambda}|}{\epsilon|\lambda|}\xrightarrow[\epsilon\longrightarrow 0]{}\frac{|||p|||_2\|\Lambda(\lambda)\|_2}{|\lambda|\cdot|p'(\lambda)|}.$$

For this, set

$$r(z):=\sum_{k=0}^{n-1}|\lambda|^k e^{-ik\theta}z^k,\quad \text{with } \theta:=\arg(\lambda),$$

which is a polynomial of degree $n-1$. Note that

(a) $r(\lambda)=|r(\lambda)|=\|\Lambda(\lambda)\|_2^2$, and

(b) $\|r\|_2=\|\Lambda(\lambda)\|_2$.

Now, let $\widetilde{\lambda}\in\mathbb{C}$ such that

$$\frac{|p(\widetilde{\lambda})|}{|r(\widetilde{\lambda})|}=\epsilon\frac{|||p|||_2}{\|\Lambda(\lambda)\|_2},$$

for a given ϵ small enough. This $\widetilde{\lambda}$ always exists, since $|p(z)|$ is arbitrarily small around the root λ. Then, set

$$\widetilde{p}(z)=p(z)-\frac{p(\widetilde{\lambda})}{r(\widetilde{\lambda})}r(z),$$

which is a monic polynomial of degree n. These $\widetilde{\lambda}$ and $\widetilde{p}(z)$ satisfy conditions (i)–(iii):

(i) It is clear by definition of $\widetilde{p}(z)$.

(ii) From $\widetilde{p}(z)-p(z)=-(p(\widetilde{\lambda})/r(\widetilde{\lambda}))r(z)$, we get

$$\|\widetilde{p}-p\|_2=|||\widetilde{p}-p|||_2=\frac{|p(\widetilde{\lambda})|}{|r(\widetilde{\lambda})|}\|r\|_2=\epsilon\frac{|||p|||_2}{\|\Lambda(\lambda)\|_2}\|r\|_2=\epsilon|||p|||_2,$$

where in the last equality we have used the property (b) of $r(z)$.

(iii) As we have seen in the first part of the proof, since $\widetilde{\lambda}$ is a root of $\widetilde{p}(z)$ and λ is a root of $p(z)$, we have

$$\frac{|\lambda-\widetilde{\lambda}|}{\epsilon|\lambda|}=\frac{\widetilde{p}(\lambda)-p(\lambda)}{\epsilon|\lambda|\cdot|p'(\lambda)|}+O(\epsilon)=\frac{(|p(\widetilde{\lambda})|/|r(\widetilde{\lambda})|)\cdot|r(\lambda)|}{\epsilon|\lambda|\cdot|p'(\lambda)|}+O(\epsilon)$$
$$=\frac{(\epsilon|||p|||_2/\|\Lambda(\lambda)\|_2)\|\Lambda(\lambda)\|_2^2}{\epsilon|\lambda|\cdot|p'(\lambda)|}+O(\epsilon)=\frac{|||p|||_2\|\Lambda(\lambda)\|_2}{|\lambda|\cdot|p'(\lambda)|}+O(\epsilon)$$

$$\xrightarrow[\epsilon \longrightarrow 0]{} \frac{\|\|p\|\|_2 \|\Lambda(\lambda)\|_2}{|\lambda| \cdot |p'(\lambda)|},$$

where in the third equality we have used the property (a) of $r(z)$.

□

8.2 Explicit formulas for the eigenvectors of Fiedler matrices

In Theorem 8.2, we present explicit expressions, in terms of λ and the coefficients of $p(z)$, of the right and left eigenvectors of any Fiedler matrix associated with an eigenvalue λ (see Remark 1.3 for our convention about eigenvectors). These expressions are already known (see [45], for example), although we present here a new proof that employs the expressions for $\mathrm{adj}(zI - M_\sigma)$ in Chapter 5. Theorem 8.2 relates the right and left eigenvectors of M_σ associated with an eigenvalue λ with the vectors x_σ, y_σ, v_σ and w_σ defined in Theorem 5.3. For convenience, in Theorem 8.2 we write explicitly the dependence on z of the vectors x_σ, y_σ, v_σ and w_σ as $x_\sigma(z)$, $y_\sigma(z)$, $v_\sigma(z)$ and $w_\sigma(z)$. Note that, since any Fiedler matrix M_σ is a *non-derogatory matrix* [1], the right and left eigenvectors of M_σ associated with an eigenvalue λ are, up to a multiplicative factor, unique.

Theorem 8.2. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let $x_\sigma(z)$, $y_\sigma(z)$, $v_\sigma(z)$ and $w_\sigma(z)$ be the vectors defined in Theorem* 5.3*. If λ is a root of $p(z)$, then*

$$r_\sigma := x_\sigma(\lambda) \quad \textit{and} \quad l_\sigma := y_\sigma(\lambda),$$

are the right and left eigenvectors of M_σ, respectively, associated with λ. Moreover, if λ is nonzero, then $r_\sigma = v_\sigma(\lambda)$ and $l_\sigma = w_\sigma(\lambda)$.

Proof. From $\mathrm{adj}(zI - M_\sigma)(zI - M_\sigma) = (zI - M_\sigma)\mathrm{adj}(zI - M_\sigma) = p(z)I$ (see [66, Ch. 4, §4]), it follows that any nonzero row of $\mathrm{adj}(\lambda I - M_\sigma)$ is the transpose of a left eigenvector of M_σ associated with the eigenvalue λ, and any nonzero column of $\mathrm{adj}(\lambda I - M_\sigma)$ is a right eigenvector of M_σ associated with the eigenvalue λ. Then, using $p(\lambda) = 0$ and (5.3), we get that $\mathrm{adj}(\lambda I - M_\sigma) = x_\sigma(\lambda) y_\sigma^T(\lambda) = r_\sigma l_\sigma^T$. Finally, from Lemma 5.9 we know that both $r_\sigma = x_\sigma(\lambda)$ and $l_\sigma = y_\sigma(\lambda)$ have an entry identically equal to one, and, therefore, the matrix $\mathrm{adj}(\lambda I - M_\sigma)$ has a nonzero column equal to r_σ and a nonzero row equal to l_σ^T.

The fact that r_σ and l_σ are equal to $v_\sigma(\lambda)$ and $w_\sigma(\lambda)$, respectively, when $\lambda \neq 0$, follows from the expressions for the entries of the vectors $x_\sigma(z)$, $y_\sigma(z)$, $v_\sigma(z)$, and $w_\sigma(z)$ in Theorem 5.3 together with the following relation between the Horner shifts of $p(z)$ and the Horner shifts of the reversal polynomial $p^{\mathrm{rev}}(z)$ of $p(z)$:

$$p_{k-1}(\lambda) = -\lambda^{-1} p^{\mathrm{rev}}_{n-k}(\lambda^{-1}),$$

for $k = 1, 2, \ldots, n$, which may be easily checked. □

Theorem 8.2, together with the expressions for $x_\sigma(z)$, $y_\sigma(z)$, $v_\sigma(z)$ and $w_\sigma(z)$ in Theorem 5.3, allows us to easily get explicit expressions for the right and left eigenvectors of any Fiedler matrix of $p(z)$ associated with an eigenvalue λ. These expressions depend on the eigenvalue λ and the Horner shifts of $p(z)$ evaluated at λ. To illustrate Theorem 8.2 we provide the following examples.

[1] The first and second Frobenius companion matrices, C_1 and C_2, of a monic polynomial $p(z)$ are non-derogatory matrices [87], that is, the geometric multiplicity of each eigenvalue is equal to one. Since Fiedler matrices of $p(z)$ are similar to each other, and since the geometric multiplicities of eigenvalues do not change under similarity, all Fiedler matrices are non-derogatory matrices.

Example 8.3. Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let C_1 and C_2 be the first and second Frobenius companion matrices of $p(z)$, respectively, and let λ be a root of $p(z)$. Recall from Section 2.2 that C_1 is a Fiedler matrix M_{σ_1} associated with any bijection σ_1 such that $\mathrm{PCIS}(\sigma_1) = (0, 0, \ldots, 0)$ and that C_2 is a Fiedler matrix M_{σ_2} associated with any bijection σ_2 such that $\mathrm{PCIS}(\sigma_2) = (1, 1, \ldots, 1)$. Then, from Theorem 8.2 we get the following expressions for the right and left eigenvectors of C_1 and C_2 associated with λ:

$$r_{\sigma_1} = l_{\sigma_2} = \begin{bmatrix} \lambda^{n-1} & \cdots & \lambda & 1 \end{bmatrix}^T \quad \text{and} \quad l_{\sigma_1} = r_{\sigma_2} = \begin{bmatrix} p_0(\lambda) & p_1(\lambda) & \cdots & p_{n-1}(\lambda) \end{bmatrix}^T.$$

Remark 8.4. Note that the right and left eigenvectors of C_1 (resp. the left and right eigenvectors of C_2) associated with an eigenvalue λ are equal to the vectors $\Lambda(z)$ and $\Pi(z)$, in (8.3) and (1.22), evaluated at $z = \lambda$.

Example 8.5. Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n > 2$, let F be the Fiedler matrix of $p(z)$ in (2.8), and let λ be a root of $p(z)$. Recall from Section 2.2 that F is a Fiedler matrix M_σ associated with any bijection σ such that $\mathrm{PCIS}(\sigma) = (0, 1, \ldots, 1)$. Then, from Theorem 8.2 we get the following expressions for the right and left eigenvectors of F associated with λ:

$$r_\sigma = \begin{bmatrix} \lambda p_0(\lambda) & \lambda p_1(\lambda) & \cdots & \lambda p_{n-2}(\lambda) & 1 \end{bmatrix}^T \quad \text{and} \quad l_\sigma = \begin{bmatrix} \lambda^{n-2} & \cdots & \lambda & 1 & p_{n-1}(\lambda) \end{bmatrix}^T.$$

Example 8.6. Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n > 2$, let P_1 be the Fiedler matrix of $p(z)$ in (2.7), and let λ be a root of $p(z)$. Recall from Section 2.2 that P_1 is a Fiedler matrix M_σ associated with any bijection σ such that $\mathrm{PCIS}(\sigma) = (1, 0, 1, 0, 1, 0, \ldots)$, then, from Theorem 8.2 we get the following expressions for the right and left eigenvectors of P_1 associated with λ:

$$r_\sigma = \left[\lambda^{\frac{n-2}{2}} p_0(\lambda) \quad \lambda^{\frac{n-2}{2}} p_1(\lambda) \quad \lambda^{\frac{n-4}{2}} \quad \lambda^{\frac{n-4}{2}} p_3(\lambda) \quad \cdots \quad 1 \quad p_{n-1}(\lambda) \right]^T \quad \text{and}$$
$$l_\sigma = \left[\lambda^{\frac{n}{2}} \quad \lambda^{\frac{n-2}{2}} \quad \lambda^{\frac{n-2}{2}} p_2(\lambda) \quad \cdots \quad \lambda \quad \lambda p_{n-2}(\lambda) \quad 1 \right]^T,$$

if n is even, and

$$r_\sigma = \left[\lambda^{\frac{n-1}{2}} \quad \lambda^{\frac{n-3}{2}} \quad \lambda^{\frac{n-3}{2}} p_2(\lambda) \quad \lambda^{\frac{n-5}{2}} \quad \lambda^{\frac{n-5}{2}} p_4(\lambda) \quad \cdots \quad 1 \quad p_{n-1}(\lambda) \right]^T \quad \text{and}$$
$$l_\sigma = \left[\lambda^{\frac{n-1}{2}} \quad \lambda^{\frac{n-1}{2}} p_1(\lambda) \quad \lambda^{\frac{n-3}{2}}, \quad \lambda^{\frac{n-3}{2}} p_3(\lambda) \quad \cdots \quad \lambda \quad \lambda p_{n-2}(\lambda) \quad 1 \right]^T,$$

if n is odd.

8.3 Eigenvalue condition numbers of Fiedler matrices

Given a matrix $A \in \mathbb{C}^{n \times n}$ and a simple nonzero eigenvalue λ of A, recall that the condition number of λ defined in (1.21) is equal to

$$\kappa(\lambda, A) = \frac{\|x\|_2 \|y\|_2}{|y^T x|} \frac{\|A\|_2}{|\lambda|}, \tag{8.5}$$

where $x, y \in \mathbb{C}^n$ are the right and left eigenvectors of A, respectively, associated with λ.

As a direct consequence of (8.5) and Theorem 8.2, we get for any Fiedler matrix M_σ an explicit formula for the condition number $\kappa(\lambda, M_\sigma)$.

Corollary 8.7. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler companion matrix of $p(z)$ associated to the bijection σ. If λ is a simple nonzero eigenvalue of M_σ, then*

$$\kappa(\lambda, M_\sigma) = \frac{\|r_\sigma\|_2 \ \|l_\sigma\|_2}{|p'(\lambda)|} \frac{\|M_\sigma\|_2}{|\lambda|}, \tag{8.6}$$

where the vectors $r_\sigma, l_\sigma \in \mathbb{C}^n$ are defined in Theorem 8.2.

Proof. From (8.5) and Theorem 8.2 we get

$$\kappa(\lambda, M_\sigma) = \frac{\|r_\sigma\|_2 \ \|l_\sigma\|_2}{|l_\sigma^T r_\sigma|} \frac{\|M_\sigma\|_2}{|\lambda|}.$$

So, we just need to check that $r_\sigma^T l_\sigma = p'(\lambda)$. Indeed, from Theorem 8.2 we get

$$\begin{aligned} l_\sigma^T r_\sigma &= \sum_{k=1}^{n} l_\sigma(k) r_\sigma(k) = \sum_{k=1}^{n} \lambda^{\mathfrak{i}_\sigma(0:n-k-1)+\mathfrak{c}_\sigma(0:n-k-1)} p_{k-1}(\lambda) \\ &= \sum_{k=1}^{n} \lambda^{n-k} p_{k-1}(\lambda) = \sum_{k=1}^{n} k a_k \lambda^k = p'(\lambda), \end{aligned}$$

where $a_n = 1$, and where we have used that $\mathfrak{i}_\sigma(0 : n-k-1) + \mathfrak{c}_\sigma(0 : n-k-1) = n-k$. □

Remark 8.8. Recall from Chapter 4 that the 2-norm $\|M_\sigma\|_2$ is not known except for the case $M_\sigma = C_1, C_2$. Nevertheless, in Chapter 3 we have explicit expressions for the 1-,∞- and Frobenius norms of any Fiedler matrix M_σ, and, since the 2-norm is equivalent to any of those norms, there exist constants c_n and $\widetilde{c}_n$ (that only depend on n) such that $c_n \|M_\sigma\|_F \leq \|M_\sigma\|_2 \leq \widehat{c}_n \|M_\sigma\|_F$ [87, p. 314].

As a particular case of Corollary 8.7, if C denotes the first or the second Frobenius companion matrix of the polynomial (1.1), we recover the expression for $\kappa(\lambda, C)$ given in [150]:

$$\kappa(\lambda, C) = \frac{\|C\|_2}{|\lambda|} \frac{\|\Lambda(\lambda)\|_2 \|\Pi(\lambda)\|_2}{|p'(\lambda)|}, \tag{8.7}$$

where the vectors $\Lambda(z)$ and $\Pi(z)$ are defined in (8.3) and (1.22), respectively.

8.4 Comparing condition numbers

Ideally, given a monic polynomial $p(z)$, for solving the root-finding problem for $p(z)$ by using a backward stable eigenvalue algorithm on a Fiedler companion matrix of $p(z)$, one would like the eigenvalues of the Fiedler matrix to be as well conditioned as the roots of the original polynomial. Since forward errors of computed eigenvalues are bounded by the backward errors times the eigenvalue condition numbers, and since Fiedler matrices have a norm equal (up to dimensional constants) to the norm of the polynomial, this would imply that the roots of $p(z)$ are computed with the forward errors expected from the sensitivity of the original data, i.e., from $p(z)$. In other words, one would like

$$\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} = O(1),$$

where, from (8.6) and (8.4), this ratio is equal to

$$\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} = \frac{\|M_\sigma\|_2}{|||p|||_2} \frac{\|r_\sigma\|_2 \|l_\sigma\|_2}{\|\Lambda(\lambda)\|_2}, \tag{8.8}$$

where r_σ, l_σ are the right and left eigenvectors of M_σ associated with the eigenvalue λ (see Theorem 8.2). In particular, if C denotes the first or the second Frobenius companion matrix of $p(z)$,

$$\frac{\kappa(\lambda, C)}{\kappa(\lambda, p)} = \frac{\|C\|_2}{|||p|||_2}\|\Pi(\lambda)\|_2, \tag{8.9}$$

where the vector $\Pi(z)$ is defined in (1.22).

The ratio of condition numbers (8.8) is a complicated function of λ and the coefficients of the polynomial $p(z)$. Theorem 8.13 provides simple upper and lower bounds for this ratio in terms of the absolute value of the coefficients of $p(z)$. To prove that the bounds in Theorem 8.13 hold we need Lemmas 8.9 and 8.10. Lemma 8.9 gives a simple upper bound for the absolute value of any Horner shift of a monic polynomial $p(z)$ evaluated at a root of $p(z)$.

Lemma 8.9. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\lambda \in \mathbb{C}$ be a root of $p(z)$, and let $\{p_k(z)\}_{k=0}^{n-1}$ be the Horner shifts of $p(z)$. Then,*

$$|p_k(\lambda)| \leq \sqrt{n}\|p\|_2,$$

for $k = 0, 1, \ldots, n-1$.

Proof. First, suppose that $|\lambda| \leq 1$. Then, from $|p_k(\lambda)| = |\lambda^k + a_{n-1}\lambda^{k-1} + \cdots + a_{n-k+1}\lambda + a_{n-k}| \leq 1 + |a_{n-1}| + \cdots + |a_0| \leq \sqrt{n}\|p\|_2$, we get the result. Second, suppose that $|\lambda| \geq 1$. Recall that the Horner shifts of $p(z)$ satisfy $p_k(z) = \lambda p_{k-1}(z) + a_{n-k}$, for $k = 1, 2, \ldots, n-1$, where $p_0(\lambda) = 1$. Since $p(\lambda) = \lambda p_{n-1}(\lambda) + a_0 = 0$, we have that $p_{n-1}(\lambda) = -a_0/\lambda$. With the previous equation, the recurrence relation $p_{k-1}(\lambda) = p_k(\lambda)/\lambda - a_{n-k}/\lambda$, for $k = 1, 2, \ldots, n-1$, implies that $p_k(\lambda) = -a_0/\lambda^{n-k} - a_1/\lambda^{n-k-1} - \cdots - a_{n-k-1}/\lambda$. Then, from $|p_k(\lambda)| = |a_0/\lambda^{n-k} + a_1/\lambda^{n-k-1} + \cdots + a_{n-k-1}/\lambda| \leq 1 + |a_{n-1}| + \cdots + |a_0| \leq \sqrt{n}\|p\|_2$, we get the result. □

Lemma 8.10 shows that Fiedler matrices associated with a polynomial $p(z)$ have a norm equal, up to dimensional constants, to the norm of the polynomial $p(z)$, and it also gives lower and upper bounds for the ratios $\|C\|_2/|||p|||_2$ and $\|M_\sigma\|_2/|||p|||_2$ in terms of the coefficients of the monic polynomial $p(z)$.

Lemma 8.10. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n > 2$, let C be the first or the second Frobenius companion matrix of $p(z)$, let M_σ be a Fiedler matrix of $p(z)$ other than the Frobenius ones, and let $\rho(p)$ be defined as*

$$\rho(p) := \sqrt{1 + \frac{1}{\max_{0 \leq k \leq n-1} |a_k|^2}}. \tag{8.10}$$

Then,

$$\frac{1}{\sqrt{2}} \leq \frac{\|C\|_2}{|||p|||_2} \leq \rho(p), \quad \frac{1}{\sqrt{n}} \leq \frac{\|M_\sigma\|_2}{|||p|||_2} \leq \sqrt{n}\rho(p), \quad \frac{1}{\sqrt{2}} \leq \frac{\|C\|_2}{\|p\|_2} \leq 1, \quad \textit{and} \quad \frac{1}{\sqrt{n}} \leq \frac{\|M_\sigma\|_2}{\|p\|_2} \leq \sqrt{n}.$$

Proof. The lower and upper bounds for $\|C\|_2/|||p|||_2$ and $\|C\|_2/\|p\|_2$ are immediate consequences of the formula for $\|C\|_2$ in (1.29). Also, using that, for any matrix $A \in \mathbb{C}^{n\times n}$, $\|A\|_F \geq \|A\|_2 \geq n^{-1/2}\|A\|_F$ [87, pp. 314], the lower and upper bounds for $\|M_\sigma\|_2/|||p|||_2$ and $\|M_\sigma\|_2/\|p\|_2$ follows from the formula for $\|M_\sigma\|_F$ in (3.3). □

Remark 8.11. Notice that to bound $\|M_\sigma\|_2/|||p|||_2$ and $\|M_\sigma\|_2/\|p\|_2$ we have used the Frobenius norm $\|M_\sigma\|_F$ instead of $\|M_\sigma\|_2$ because explicit expressions for the 2-norm of Fiedler matrices are not known when $M_\sigma \neq C_1, C_2$ (see Chapter 4).

Remark 8.12. Notice that $\rho(p)$ in (8.10) is always equal or greater than one, and that there are polynomials for which $\rho(p)$, and therefore the upper bounds for $\|C\|_2/|||p|||_2$ and $\|M_\sigma\|_2/|||p|||_2$ in Lemma 8.10, may be as large as desired.

Theorem 8.13. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. If λ is a simple nonzero root of $p(z)$, then*

$$\frac{1}{\sqrt{2}} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq n\rho(p)\|p\|_2 \tag{8.11}$$

if $M_\sigma = C_1, C_2$, and

$$\frac{1}{n} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq n^{5/2}\rho(p)\|p\|_2^2 \tag{8.12}$$

if $M_\sigma \neq C_1, C_2$, where $\rho(p)$ is defined in (8.10)*.*

Proof. We prove first (8.11), that is, we have to bound (8.9). Since $\kappa(\lambda, C_1) = \kappa(\lambda, C_2)$ we will focus only on C_1. From Example 8.3 and Lemma 8.9, we have that the modulus of all the entries of $l_\sigma = \Pi(\lambda)$ are bounded by $\sqrt{n}\|p\|_2$, and, therefore, $\|l_\sigma\|_2 = \|\Pi(\lambda)\|_2 \leq n\|p\|_2$. With the previous inequality, the upper bound in (8.11) follows from Lemma 8.10. Finally, the lower bound in (8.11) follows from Lemma 8.10 and from $\|\Pi(\lambda)\|_2 \geq 1$, since $p_0(\lambda) = 1$.

Next, we prove (8.12), that is, we have to bound (8.8) when $M_\sigma \neq C_1, C_2$. From Lemma 8.10 we have that $\|M_\sigma\|_2/|||p|||_2 \leq \sqrt{n}\rho(p)$. So, to prove that the upper bound in (8.12) holds, we need to show that $\|r_\sigma\|_2\|l_\sigma\|_2/\|\Lambda(\lambda)\|_2 \leq n^2\|p\|_2^2$. In order to do this, we have to distinguish two cases: $|\lambda| \leq 1$ and $|\lambda| > 1$.

When $|\lambda| \leq 1$, from Theorems 5.3 and 8.2 it follows that, for $k = 1, 2, \ldots, n$, the modulus of the kth entry of r_σ and of l_σ is bounded by $\max\{1, |p_{k-1}(\lambda)|\}$, so, using Lemma 8.9 we get that the modulus of these entries are bounded by $\sqrt{n}\|p\|_2$, and, therefore, $\|r_\sigma\|_2\|l_\sigma\|_2 \leq n^2\|p\|_2^2$. With the previous inequality and using that $\|\Lambda(\lambda)\|_2 \geq 1$, the result follows.

When $|\lambda| > 1$, using $\|\Lambda(\lambda)\|_2 \geq |\lambda^{n-1}|$ and $n - 1 = \mathfrak{i}_\sigma(0 : n-2) + \mathfrak{c}_\sigma(0 : n-2)$, we get

$$\frac{\|r_\sigma\|_2\|l_\sigma\|_2}{\|\Lambda(\lambda)\|_2} \leq \frac{\|r_\sigma\|_2\|l_\sigma\|_2}{|\lambda^{n-1}|} = \frac{\|r_\sigma\|_2}{|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|}\frac{\|l_\sigma\|_2}{|\lambda^{\mathfrak{c}_\sigma(0:n-2)}|}.$$

So, we need to bound the modulus of the entries of $r_\sigma/\lambda^{\mathfrak{i}_\sigma(0:n-2)}$ and $l_\sigma/\lambda^{\mathfrak{c}_\sigma(0:n-2)}$. Since $|\lambda| \geq 1$, and, for $k = 1, 2, \ldots, n$, $\mathfrak{i}_\sigma(0 : n-2) \geq \mathfrak{i}_\sigma(0 : n-k-1)$ and $\mathfrak{c}_\sigma(0 : n-2) \geq \mathfrak{c}_\sigma(0 : n-k-1)$, from Theorems 5.3 and 8.2 it follows that the modulus of the kth entry of $r_\sigma/\lambda^{\mathfrak{i}_\sigma(0:n-2)}$ and of $l_\sigma/\lambda^{\mathfrak{c}_\sigma(0:n-2)}$ is bounded by $\max\{1, |p_{k-1}(\lambda)|\}$, so, using Lemma 8.9 we get that the modulus of these entries are bounded by $\sqrt{n}\|p\|_2$, and, therefore, $\|r_\sigma\|_2\|l_\sigma\|_2/\|\Lambda(\lambda)\|_2 \leq n^2\|p\|_2^2$. With the last inequality, the result follows.

Finally, we prove that the lower bound in (8.12) holds. From Lemma 8.10, we have that $\|M_\sigma\|_2/|||p|||_2 \geq 1/\sqrt{n}$, so, we only need to show that $\|r_\sigma\|_2\|l_\sigma\|_2/\|\Lambda(\lambda)\|_2 \geq 1/\sqrt{n}$. In order to prove the previous inequality, first, notice that $\|r_\sigma\|_2\|l_\sigma\| \geq \max\{1, |r_\sigma(1)l_\sigma(1)|\} = \max\{1, |\lambda|^{n-1}\}$, where we have used that the vectors r_σ and l_σ have, at least, an entry equal to 1 (see Lemma 5.9), and that $r_\sigma(1)l_\sigma(1) = \lambda^{n-1}$ (see Theorems 5.3 and 8.2). Also notice that $\|\Lambda(\lambda)\|_2 = \sqrt{\sum_{i=0}^{n-1} |\lambda|^{2i}} \leq \sqrt{n}\max\{1, |\lambda|^{n-1}\}$. Then, from the two previous observation, it follows that $\|r_\sigma\|_2\|l_\sigma\|_2/\|\Lambda(\lambda)\|_2 \geq 1/\sqrt{n}$.

□

Remark 8.14. Probably, if explicit expressions for the 2-norm of Fiedler matrices other than the Frobenius ones were available, upper and lower bounds sharper than the ones in (8.12) could be found. Although, in Example 8.15, we will show that the presence of $\|p\|_2^2$ is necessary in the upper bound in (8.12) .

The presence of $\|p\|_2$ and $\rho(p)$ in the upper bounds in (8.11) and (8.12) shows that these bounds are large if and only if $\max\{|a_{n-1}|,\dots,|a_1|,|a_0|\}$ is either large or close to zero. Therefore, from Theorem 8.13, we get the following conclusions:

(C1) When $\max\{|a_{n-1}|,\dots,|a_1|,|a_0|\}$ is moderate and not close to zero, that is, of order $\Theta(1)$, then the eigenvalues of Fiedler matrices are as well conditioned as the roots of the monic polynomial $p(z)$. In this case, from the point of view of condition numbers, any Fiedler matrix can be used for solving the root-finding problem for $p(z)$.

(C2) When $\max\{|a_{n-1}|,\dots,|a_1|,|a_0|\}$ is large or close to zero, the eigenvalues of any Fiedler companion matrix may be potentially more ill conditioned than the roots of the monic polynomial $p(z)$.

It is evident that there exist polynomials for which the upper bounds in (8.11) and (8.12) can be as large as desired, but this does not necessarily mean that the ratio (8.8) is large for these polynomials. Nevertheless, Example 8.15 shows that there exist polynomials for which a large upper bound in (8.11) or in (8.12) implies that the ratio (8.8) is also large.

Example 8.15. Let $A \geq 2$, and consider the monic polynomial $p(z) = z^n + (A-1)z^{n-1} - A$. It may be checked that $\lambda = 1$ is a root of $p(z)$, and that the Horner shifts of $p(z)$, evaluated at $z = 1$, satisfy $p_0(1) = 1$, and, for $k = 1, 2, \dots, n-1$, $p_k(1) = A$. Now, given a Fiedler matrix M_σ of $p(z)$, we are going to prove that, for $\lambda = 1$, the ratio between (8.8) and the upper bounds in Theorem 8.13 is larger than a quantity that depends only on n.

First, suppose that $M_\sigma = C_1, C_2$. From (8.9) and Lemma 8.10 we get

$$\frac{\kappa(1, M_\sigma)}{\kappa(1,p)} = \frac{\|M_\sigma\|_2}{|||p|||_2} \| \begin{bmatrix} 1 & A & A & \cdots & A \end{bmatrix}^T \|_2 \geq \sqrt{\frac{n-1}{2}} A.$$

Also, the upper bound in (8.11) is equal to $n(1+1/A^2)^{1/2} \| \begin{bmatrix} 1 & A-1 & 0 & \cdots & 0 & -A \end{bmatrix}^T \|_2$ wich is less than or equal to $\sqrt{6}nA$. From this, we get that the ratio between (8.9) and the upper bound in (8.11) is larger than or equal to $(n-1)^{1/2}/(2\sqrt{3}n)$. So, taking A and n large enough, we have a polynomial $p(z)$ for which a large upper bound in (8.11) implies that the ratio (8.9) is large.

Second, suppose that $M_\sigma \neq C_1, C_2$. Observe that, since M_σ is not one of the Frobenius companion matrices, if $\mathrm{PCIS}(\sigma) = (v_0, v_1, \dots, v_{n-2})$, then there exist $i, j \in \{2, 3, \dots, n\}$ such that $v_{n-i} \neq v_{n-j}$. Suppose that $v_{n-i} = 1$ and $v_{n-j} = 0$ (when $v_{n-i} = 0$ and $v_{n-j} = 1$ the argument is similar, so we omit it). If $v_{n-i} = 1$ and $v_{n-j} = 0$, and if r_σ and l_σ are the right and left eigenvectors of M_σ associated with $\lambda = 1$, respectively, then Theorem 8.2 implies $\|r_\sigma\|_2 \|l_\sigma\|_2 \geq |r_\sigma(i)| \cdot |l_\sigma(j)| = |p_{i-1}(1)| \cdot |p_{j-1}(1)| = A^2$, and, therefore,

$$\frac{\kappa(1, M_\sigma)}{\kappa(1,p)} = \frac{\|M_\sigma\|_2}{|||p|||_2} \frac{\|r_\sigma\|_2 \|l_\sigma\|_2}{\|\Lambda(1)\|_2} \geq \frac{A^2}{n},$$

where we have used Lemma 8.10 and $\|\Lambda(1)\|_2 = n^{1/2}$. Also, the upper bound in (8.12) is equal to $n^{5/2}(1+1/A^2)^{1/2} \| \begin{bmatrix} 1 & A-1 & 0 & \cdots & 0 & -A \end{bmatrix}^T \|_2^2$ which is less than or equal to $3\sqrt{2}n^{5/2}A^2$. From this, we get that the ratio between (8.8) and the upper bound in (8.12) is larger than or equal to $1/(3\sqrt{2}n^{7/2})$. So, again, taking A and n large enough, we have a polynomial $p(z)$ for which a large upper bound in (8.12) implies that the ratio (8.8) is large.

Notice that the upper bound in (8.12) is larger than the upper bound in (8.11). This suggests that the eigenvalues of Fiedler companion matrices other than the Frobenius ones may be potentially more ill conditioned than the eigenvalues of the Frobenius companion matrices. Since in the

polynomial root-finding problem using Fiedler companion matrices is important to know whether or not the eigenvalue condition numbers of Fiedler companion matrices other than the Frobenius ones are much larger (or smaller) that the eigenvalue condition numbers of Frobenius companion matrices, it is of fundamental importance to study the ratio

$$\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} = \frac{\|M_\sigma\|_2}{\|C\|_2} \frac{\|r_\sigma\|_2 \|l_\sigma\|_2}{\|\Lambda(\lambda)\|_2 \|\Pi(\lambda)\|_2}, \tag{8.13}$$

where $C = C_1, C_2$ and $M_\sigma \neq C_1, C_2$, that we obtain from (8.6) and (8.7). In Theorem 8.17 we give upper bounds of (8.13) in terms of the norm of the vector $\Pi(\lambda)$, and, also, in terms of the absolute values of the coefficients of $p(z)$. Lemma 8.16 will be useful in establishing these bounds.

Lemma 8.16. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let C be the first or the second Frobenius companion matrix, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. Then, $n^{-1/2}\|M_\sigma\|_2 \leq \|C\|_2 \leq n^{1/2}\|M_\sigma\|_2$.*

Proof. The result follows from $n^{-1/2}\|A\|_F \leq \|A\|_2 \leq \|A\|_F$, for any matrix $A \in \mathbb{C}^{n\times n}$, [87, pp. 314] and $\|M_\sigma\|_F = \|C\|_F$ (see Corollary 3.5). □

Theorem 8.17. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n > 2$, let C denote the first or the second Frobenius companion matrix of $p(z)$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection such that PCIS$(\sigma) \neq (0, \ldots, 0), (1, \ldots, 1)$, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. If λ is a simple nonzero root of $p(z)$, then,*

$$1 \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} \leq n^{3/2}\|\Pi(\lambda)\|_2 \leq n^{5/2}\|p\|_2, \tag{8.14}$$

if $\kappa(\lambda, M_\sigma) \geq \kappa(\lambda, C)$, and

$$1 \leq \frac{\kappa(\lambda, C)}{\kappa(\lambda, M_\sigma)} \leq n\|\Pi(\lambda)\|_2 \leq n^2\|p\|_2, \tag{8.15}$$

if $\kappa(\lambda, C) \geq \kappa(\lambda, M_\sigma)$, where $\Pi(\lambda)$ is the vector defined in (1.22).

Proof. First, we prove (8.14). From (8.13) and Lemma 8.16, we get

$$1 \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} \leq n^{1/2} \frac{\|r_\sigma\|_2 \|l_\sigma\|_2}{\|\Lambda(\lambda)\|_2 \|\Pi(\lambda)\|_2},$$

Therefore, to get the first inequality in (8.14) we need to check that $\|r_\sigma\|_2\|l_\sigma\|_2/(\|\Lambda(\lambda)\|_2\|\Pi(\lambda)\|_2) \leq n\|\Pi(\lambda)\|_2$ holds. In order to do this we distinguish two cases: $|\lambda| \leq 1$ and $|\lambda| > 1$.

If $|\lambda| \leq 1$, using $\|\Lambda(\lambda)\|_2 \geq 1$, we get

$$\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} \leq n^{1/2} \frac{\|r_\sigma\|_2 \|l_\sigma\|_2}{\|\Pi(\lambda)\|_2}.$$

So, we need to bound the norm of the vectors $r_\sigma/\|\Pi(\lambda)\|_2$ and l_σ. For $k = 1, 2, \ldots, n$, Theorem 8.2 implies

$$\frac{|r_\sigma(k)|}{\|\Pi(\lambda)\|_2} = \frac{|\lambda^{\mathrm{i}_\sigma(0:n-k-1)} p_{k-1}(\lambda)|}{\|\Pi(\lambda)\|_2} \leq \frac{|p_{k-1}(\lambda)|}{\|\Pi(\lambda)\|_2} \leq 1,$$

if $v_{n-k} = 1$, or

$$\frac{|r_\sigma(k)|}{\|\Pi(\lambda)\|_2} = \frac{|\lambda^{\mathrm{i}_\sigma(0:n-k-1)}|}{\|\Pi(\lambda)\|_2} \leq \frac{1}{\|\Pi(\lambda)\|_2} \leq 1,$$

if $v_{n-k}=0$, and

$$|l_\sigma(k)| = |\lambda^{\mathfrak{c}_\sigma(0:n-k-1)}p_{k-1}(\lambda)| \leq |p_{k-1}(\lambda)| \leq \|\Pi(\lambda)\|_2,$$

if $v_{n-k}=0$, or

$$|l_\sigma(k)| = |\lambda^{\mathfrak{c}_\sigma(0:n-k-1)}| \leq 1 \leq \|\Pi(\lambda)\|_2,$$

if $v_{n-k}=1$, where we have used that $\|\Pi(\lambda)\|_2 \geq |p_{k-1}(\lambda)|$, for $k=1,2,\ldots,n$, (in particular, for $k=1$, we have that $\|\Pi(\lambda)\|_2 \geq 1$, since $p_0(\lambda)=1$). From the previous bounds, we get $\|r_\sigma\|_2/\|\Pi(\lambda)\|_2 \leq n^{1/2}$ and $\|l_\sigma\|_2 \leq n^{1/2}\|\Pi(\lambda)\|_2$, and from this, the first bound in (8.14) follows. The rightmost bound in (8.14) follows from Lemma 8.9.

If $|\lambda|>1$, then

$$\frac{\kappa(\lambda,M_\sigma)}{\kappa(\lambda,C)} \leq n^{1/2}\frac{\|r_\sigma\|_2\|l_\sigma\|_2}{|\lambda^{n-1}|\|\Pi(\lambda)\|_2} \leq n^{1/2}\frac{\|r_\sigma\|_2}{\|\Pi(\lambda)\|_2|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|}\frac{\|l_\sigma\|_2}{|\lambda^{\mathfrak{c}_\sigma(0:n-2)}|},$$

where we have used Lemma 8.16 together with the fact that $\|\Lambda(\lambda)\|_2 \geq |\lambda^{n-1}|$, and also that $n-1 = \mathfrak{i}_\sigma(0:n-2)+\mathfrak{c}_\sigma(0:n-2)$. So, we need to bound the norm of the vectors $r_\sigma/(\|\Pi(\lambda)\|_2|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|)$ and $l_\sigma/|\lambda^{\mathfrak{c}_\sigma(0:n-2)}|$. For $k=1,2,\ldots,n$, Theorem 8.2 implies

$$\frac{|r_\sigma(k)|}{\|\Pi(\lambda)\|_2|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|} = \frac{|\lambda^{\mathfrak{i}_\sigma(0:n-k-1)}p_{k-1}(\lambda)|}{\|\Pi(\lambda)\|_2|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|} \leq \frac{|p_{k-1}(\lambda)|}{\|\Pi(\lambda)\|_2} \leq 1,$$

if $v_{n-k}=1$, or

$$\frac{|r_\sigma(k)|}{\|\Pi(\lambda)\|_2|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|} = \frac{|\lambda^{\mathfrak{i}_\sigma(0:n-k-1)}|}{\|\Pi(\lambda)\|_2|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|} \leq \frac{1}{\|\Pi(\lambda)\|_2} \leq 1,$$

if $v_{n-k}=0$, and

$$\frac{|l_\sigma(k)|}{|\lambda^{\mathfrak{c}_\sigma(0:n-2)}|} = \frac{|\lambda^{\mathfrak{c}_\sigma(0:n-k-1)}p_{k-1}(\lambda)|}{|\lambda^{\mathfrak{c}_\sigma(0:n-2)}|} \leq |p_{k-1}(\lambda)| \leq \|\Pi(\lambda)\|_2$$

if $v_{n-k}=0$, or

$$\frac{|l_\sigma(k)|}{|\lambda^{\mathfrak{c}_\sigma(0:n-2)}|} = \frac{|\lambda^{\mathfrak{c}_\sigma(0:n-k-1)}|}{|\lambda^{\mathfrak{c}_\sigma(0:n-2)}|} \leq 1 \leq \|\Pi(\lambda)\|_2,$$

if $v_{n-k}=1$, where we have used $\|\Pi(\lambda)\|_2 \geq |p_{k-1}(\lambda)|$, and $\mathfrak{i}_\sigma(0:n-2) \geq \mathfrak{i}_\sigma(0:n-k-1)$ and $\mathfrak{c}_\sigma(0:n-2) \geq \mathfrak{c}_\sigma(0:n-k-1)$, for $k=1,2,\ldots,n$. From the previous bounds, we have $\|r_\sigma\|_2/(\|\Pi(\lambda)\|_2|\lambda^{\mathfrak{i}_\sigma(0:n-2)}|) \leq n^{1/2}$ and $\|l_\sigma\|_2/|\lambda^{\mathfrak{c}_\sigma(0:n-2)}| \leq n^{1/2}\|\Pi(\lambda)\|_2$, and from this, the first bound in (8.14) follows. The rightmost bound follows from Lemma 8.9.

Next, we prove (8.15). From (8.13) and Lemma 8.16, we get

$$\frac{\kappa(\lambda,C)}{\kappa(\lambda,M_\sigma)} = \frac{\|C\|_2}{\|M_\sigma\|_2}\frac{\|\Lambda(\lambda)\|_2\|\Pi(\lambda)\|_2}{\|r_\sigma\|_2\|l_\sigma\|_2} \leq n^{1/2}\frac{\|\Lambda(\lambda)\|_2\|\Pi(\lambda)\|_2}{\|r_\sigma\|_2\|l_\sigma\|_2}$$

Therefore, to get the first inequality in (8.15) we need to check that $\|\Lambda(\lambda)\|_2\|\Pi(\lambda)\|_2/(\|r_\sigma\|_2\|l_\sigma\|_2) \leq n^{1/2}\|\Pi(\lambda)\|_2$ holds. In order to do this, again, we distinguish the cases: $|\lambda|\leq 1$ and $|\lambda|>1$.

If $|\lambda|\leq 1$ then, using $\|r_\sigma\|_2, \|l_\sigma\|_2 \geq 1$ (see Lemma 5.9), and $\|\Lambda(\lambda)\|_2 \leq n^{1/2}$ when $|\lambda|\leq 1$, we get

$$\frac{\kappa(\lambda,C)}{\kappa(\lambda,M_\sigma)} \leq n\|\Pi(\lambda)\|_2.$$

If $|\lambda|>1$, then, using $\|r_\sigma\|_2\|l_\sigma\|_2 \geq |r_\sigma(1)l_\sigma(1)| = |\lambda^{n-1}|$, and $\|\Lambda(\lambda)\|/|\lambda^{n-1}| \leq n^{1/2}$ when $|\lambda|\geq 1$, we get

$$\frac{\kappa(\lambda,C)}{\kappa(\lambda,M_\sigma)} \leq n^{1/2}\frac{\|\Lambda(\lambda)\|_2}{|\lambda^{n-1}|}\|\Pi(\lambda)\|_2 \leq n\|\Pi(\lambda)\|_2,$$

The rightmost bound in (8.15) follows from Lemma 8.9. □

The presence of $\|p\|_2$ in the rightmost upper bounds in (8.14) and (8.15) shows that these bounds are large if and only if $\max\{|a_{n-1}|,\dots,|a_1|,|a_0|\}$ is large. Therefore, from Theorem 8.17 we get the following conclusions:

(C3) From the point of view of condition numbers, when $\max\{|a_{n-1}|,\dots,|a_1|,|a_0|\}$ is moderate, that is, $O(1)$, any Fiedler matrix can be used for solving the root-finding problem for $p(z)$ with the same reliability as Frobenius companion matrices.

(C4) The ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ may be potentially large (or small) for polynomials with large coefficients. In this case, for some polynomials there can be other Fiedler matrices than the Frobenius ones which are more convenient from the point of view of conditioning, and viceversa.

It is evident that there exist polynomials for which the rightmost upper bounds in (8.14) and (8.15) can be as large as desired, but this does not imply necessarily that the ratio of eigenvalue condition numbers is also large for these polynomials. In Example 8.18, given any Fiedler matrix $M_\sigma \neq C_1, C_2$, we show that there exist polynomials such that a large upper bound in Theorem 8.17 implies a large ratio between the eigenvalue condition numbers of M_σ and C_1 or C_2. In fact, we show that for these polynomials, up to a constant that depends only on the size of the problem, the rightmost bounds in Theorem 8.17 correctly predict the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ or $\kappa(\lambda, C)/\kappa(\lambda, M_\sigma)$.

Example 8.18. We first focus on the rightmost bound in (8.14). Given $A > 2$, consider the monic polynomial $p(z) = z^n + (A-1)z^{n-1} - A$. Recall from Example 8.15 that $\lambda = 1$ is a root of $p(z)$, and that the Horner shifts of $p(z)$, evaluated at $z = 1$, satisfy $p_0(1) = 1$, and, for $k = 1, 2, \dots, n-1$, $p_k(1) = A$. If C denotes one of the Frobenius companion matrices of $p(z)$ and M_σ denotes a Fiedler matrix of $p(z)$ other than the Frobenius ones, we are going to show that the ratio between $\kappa(1, M_\sigma)/\kappa(1, C)$ and the rightmost bound in (8.14) are both larger than a quantity that depends only on n.

First, we need to bound $\kappa(1, M_\sigma)/\kappa(1, C)$. This may be done using the following inequalities. From Lemma 8.16 we get $\|M_\sigma\|_2/\|C\|_2 \geq n^{-1/2}$, also, from Example 8.15, recall that if r_σ and l_σ are the right and left eigenvectors of M_σ, respectively, associated with $\lambda = 1$, then $\|r_\sigma\|_2\|l_\sigma\|_2 \geq A^2$, and, finally, $\|\Lambda(1)\|_2\|\Pi(1)\|_2 = n^{1/2}\| \begin{bmatrix} 1 & A & \cdots & A \end{bmatrix}^T \|_2 \leq nA$. From these three inequalities we get

$$\frac{\kappa(1, M_\sigma)}{\kappa(1, C)} = \frac{\|M_\sigma\|_2}{\|C\|_2} \frac{\|r_\sigma\|_2\|l_\sigma\|_2}{\|\Lambda(1)\|_2\|\Pi(1)\|_2} \geq n^{-3/2}A.$$

Also, the rightmost bound in (8.14) is equal to $n^{5/2}\| \begin{bmatrix} 1 & A-1 & 0 & \cdots & 0 & -A \end{bmatrix}^T \|_2$, which is larger than or equal to $n^{5/2}A$ and smaller than or equal to $\sqrt{3}n^{5/2}A$. So, the ratio between $\kappa(1, M_\sigma)/\kappa(1, C)$ and the rightmost bound in (8.14) is larger than $n^{-4}/\sqrt{3}$. Therefore, taking A and n large enough, for the polynomial $p(z)$ a large rightmost bound in (8.14) implies that the ratio $\kappa(1, M_\sigma)/\kappa(1, C)$ is also large.

Next, we focus on the rightmost bound in (8.15). Given a bijection $\sigma : \{0, 1, \dots, n-1\} \to \{1, \dots, n\}$ with $\text{PCIS}(\sigma) = (v_0, v_1, \dots, v_{n-2}) \neq (0, \dots, 0), (1, \dots, 1)\}$, then, it may be checked that there are $i, j \in \{2, 3, \dots, n\}$ with $i > j$ such that $v_{n-i} = 1$ and $v_{n-j} = 0$ or $v_{n-i} = 0$ and $v_{n-j} = 1$. Suppose that $v_{n-i} = 1$ and $v_{n-j} = 0$ (if $v_{n-i} = 0$ and $v_{n-j} = 1$ the argument is completely similar, so we omit it). Let $A > 1$ and $\epsilon < 1$ such that $\epsilon A < 1$, and consider the monic polynomial $q(z) = z^n + (A-\epsilon)z^{n-1} - \epsilon A z^{n-2}$ if $j = 2$, or $q(z) = z^n - \epsilon z^{n-1} + Az^{n-j+1} - \epsilon A z^{n-j}$ if $j > 2$. It may be easily checked that $\lambda = \epsilon$ is a root of $q(z)$ and that the Horner shifts of $q(z)$ satisfy $q_0(\epsilon) = 1$, $q_{j-1}(\epsilon) = A$, and $q_{k-1}(\epsilon) = 0$ when $k \neq j$.

Now, let C denote the first or the second Frobenius companion matrix of $q(z)$, and let M_σ be the Fiedler matrix of $q(z)$ associated with the bijection σ. If r_σ and l_σ denote the right and left eigenvector of M_σ, respectively, associated with $\lambda = \epsilon$, from the values of the Horner shifts of $q(z)$ in ϵ, Theorem 8.2, and $\epsilon A < 1$ it can be proved that $\|r_\sigma\|_2\|l_\sigma\|_2 \leq n$. Hence

$$\frac{\kappa(\epsilon, C)}{\kappa(\epsilon, M_\sigma)} = \frac{\|C\|_2}{\|M_\sigma\|_2}\frac{\|\Lambda(\epsilon)\|_2\|\Pi(\epsilon)\|_2}{\|r_\sigma\|_2\|l_\sigma\|_2} \geq \frac{A}{n^{3/2}}.$$

Also the rightmost upper bound in (8.15) is less than or equal to $2n^2A$. So, the ratio between $\kappa(\epsilon, C)/\kappa(\epsilon, M_\sigma)$ and the rightmost bound in (8.15) is larger than $n^{-7/2}/2$. Therefore, taking A large enough and ϵ small enough, for the polynomial $q(z)$ a large rightmost bound in (8.15) implies a large ratio $\kappa(\epsilon, C)/\kappa(\epsilon, M_\sigma)$.

Theorem 8.17 and Example 8.18 suggest that for some polynomials one should avoid using a Fiedler matrix M_σ other than the Frobenius ones, and to use, instead, the Frobenius companion matrices. However, Theorem 8.19 shows that this could only happen for a polynomial $p(z)$ whose roots are very ill-conditioned either as eigenvalues of the Frobenius companion matrices or as eigenvalues of the Fiedler matrix M_σ compared to the conditioning of the roots of $p(z)$. In other words:

$$\kappa(\lambda, M_\sigma) \gg \kappa(\lambda, C) \longrightarrow \kappa(\lambda, M_\sigma) \gg \kappa(\lambda, p) \quad \text{and} \quad \kappa(\lambda, C) \gg \kappa(\lambda, p).$$

By contrast, if the roots of a polynomial $p(z)$ are much more ill-conditioned as eigenvalues of one of the Frobenius companion matrices than they are as eigenvalues of another Fiedler matrix M_σ, then, Theorem 8.19 implies that the roots of $p(z)$ are very ill-conditioned as eigenvalues of the Frobenius matrices compared to their conditioning as a roots of $p(z)$. In other words:

$$\kappa(\lambda, C) \gg \kappa(\lambda, M_\sigma) \longrightarrow \kappa(\lambda, C) \gg \kappa(\lambda, p).$$

But $\kappa(\lambda, C) \gg \kappa(\lambda, M_\sigma)$ does not imply that the roots of $p(z)$ are ill-conditioned as eigenvalues of M_σ compared with their conditioning as roots of $p(z)$.

Theorem 8.19. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial, let C denote the first or the second Frobenius companion matrix of $p(z)$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection such that PCIS$(\sigma) \neq (0, \ldots, 0), (1, \ldots, 1)$, and let M_σ be the Fiedler matrix of $p(z)$ associated with σ. If λ is a simple nonzero root of $p(z)$, then the following results hold.*

(a) *If $\kappa(\lambda, M_\sigma) \geq \kappa(\lambda, C)$, then*

$$\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \geq \frac{1}{\sqrt{2}}\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} \quad \textit{and} \quad \frac{\kappa(\lambda, C)}{\kappa(\lambda, p)} \geq \frac{1}{n^{3/2}\sqrt{2}}\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)}. \tag{8.16}$$

(b) *If $\kappa(\lambda, C) \geq \kappa(\lambda, M_\sigma)$, then*

$$\frac{\kappa(\lambda, C)}{\kappa(\lambda, p)} \geq \frac{1}{n\sqrt{2}}\frac{\kappa(\lambda, C)}{\kappa(\lambda, M_\sigma)}. \tag{8.17}$$

Proof. First, we prove part (a). From Lemma 8.10 and using $\|\Pi(\lambda)\|_2 \geq 1$, we have

$$\begin{aligned}\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} &= \frac{\|M_\sigma\|_2}{\|C\|_2}\frac{\|r_\sigma\|_2\|l_\sigma\|_2}{\|\Lambda(\lambda)\|_2\|\Pi(\lambda)\|_2} \leq \frac{\|M_\sigma\|_2}{\|C\|_2}\frac{\|r_\sigma\|_2\|l_\sigma\|_2}{\|\Lambda(\lambda)\|_2} \\ &= \frac{\|M_\sigma\|_2}{|\lambda|}\frac{\|r_\sigma\|_2\|l_\sigma\|_2}{|p'(\lambda)|}\frac{|\lambda|}{|||p|||_2}\frac{|p'(\lambda)|}{\|\Lambda(\lambda)\|_2}\frac{|||p|||_2}{\|C\|_2} \leq \sqrt{2}\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)}.\end{aligned}$$

Also, from Lemma 8.10 and using the upper bound in (8.14), we have

$$\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} \leq n^{3/2}\|\Pi(\lambda)\|_2 = n^{3/2}\frac{\|C\|_2}{|\lambda|}\frac{\|\Pi(\lambda)\|_2\|\Lambda(\lambda)\|_2}{|p'(\lambda)|}\frac{|\lambda|}{|||p|||_2}\frac{|p'(\lambda)|}{\|\Lambda(\lambda)\|_2}\frac{|||p|||_2}{\|C\|_2} \leq n^{3/2}\sqrt{2}\frac{\kappa(\lambda, C)}{\kappa(\lambda, p)}.$$

Next, we prove part (b). From Lemma 8.10 and using the upper bound in (8.15), we have

$$\frac{\kappa(\lambda, C)}{\kappa(\lambda, M_\sigma)} \leq n\|\Pi(\lambda)\|_2 = n\frac{\|C\|_2}{|\lambda|}\frac{\|\Pi(\lambda)\|_2\|\Lambda(\lambda)\|_2}{|p'(\lambda)|}\frac{|\lambda|}{|||p|||_2}\frac{|p'(\lambda)|}{\|\Lambda(\lambda)\|_2}\frac{|||p|||_2}{\|C\|_2} \leq n\sqrt{2}\frac{\kappa(\lambda, C)}{\kappa(\lambda, p)}.$$

□

From Theorem 8.19 together with Theorem 8.17 we get the following conclusions:

(C5) From the point of view of condition numbers, there are polynomials for which Frobenius companion matrices may be better suited than the rest of Fiedler matrices in the problem of computing their roots, but only in situations where it is not recommended to compute them neither as eigenvalues of the Frobenius companion matrices or as eigenvalues of any other Fiedler matrix.

(C6) From the point of view of condition numbers, there may be polynomials for which one should avoid computing their roots as the eigenvalues of Frobenius companion matrices and to use, instead, another Fiedler matrix. Although Theorem 8.19 does not show how to identify these polynomials and how to know in advance which Fiedler matrix might be used.

The difference between the statements of parts (a) and (b) in Theorem 8.19 is striking, but the next example shows that if the ratio $\kappa(\lambda, C)/\kappa(\lambda, M_\sigma)$ is large then the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ is not necessarily large.

Example 8.20. Consider the following polynomial of degree 3: $p(z) = z^3 - z^2(\epsilon + 1/\epsilon) + z$, with $\epsilon \ll 1$, whose roots are $\epsilon, \epsilon^{-1}, 0$. Let C denotes the first or the second Frobenius companion matrix of $p(z)$, and let M_σ be the Fiedler matrix of $p(z)$ associated with a bijection σ such that $\mathrm{PCIS}(\sigma) = (0, 1)$. It may be easily checked that

$$\frac{\kappa(\epsilon, C)}{\kappa(\epsilon, M_\sigma)} = \frac{\|C\|_2}{\|M_\sigma\|_2}\frac{\|\begin{bmatrix}\epsilon^2 & \epsilon & 1\end{bmatrix}^T\|_2\|\begin{bmatrix}1 & -\epsilon^{-1} & 0\end{bmatrix}^T\|_2}{\|\begin{bmatrix}\epsilon & 1 & 0\end{bmatrix}^T\|_2\|\begin{bmatrix}\epsilon & -1 & 1\end{bmatrix}^T\|_2} \geq \frac{1}{3\epsilon},$$

and

$$\frac{\kappa(\epsilon, M_\sigma)}{\kappa(\epsilon, p)} = \frac{\|M_\sigma\|_2}{|||p|||_2}\frac{\|\begin{bmatrix}\epsilon & 1 & 0\end{bmatrix}^T\|_2\|\begin{bmatrix}\epsilon & -1 & 1\end{bmatrix}^T\|_2}{\|\begin{bmatrix}\epsilon^2 & \epsilon & 1\end{bmatrix}^T\|_2} \leq 3,$$

where to get the first inequality we have used Lemma 8.16, and to get the last inequality we have used that $\|M_\sigma\|_2/|||p|||_2 \leq \|M_\sigma\|_F/|||p|||_2 \leq \sqrt{3}$. So, taking ϵ small enough, $\kappa(\lambda, C)/\kappa(\lambda, M_\sigma)$ may be as large as desired, but $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ is bounded by a constant that is independent of ϵ.

8.5 Pseudospectra of Fiedler matrices

In this section we establish several mathematical relationships between the pseudozero sets of a monic polynomial $p(z)$ and the pseudospectra of the associated Fiedler matrices, and we also show how to estimate accurately pseudospectra of Fiedler matrices in a fast way. These results are generalizations of the results in [150], valid only for the Frobenius companion matrices, to all Fiedler matrices.

Given the monic polynomial $p(z)$ (1.1) and a Fiedler companion matrix M_σ of $p(z)$, recall from Section 1.2.2.2 that the ϵ-pseudozero set of $p(z)$, denoted by $Z_\epsilon(p)$, is the set

$$Z_\epsilon(p) = \left\{ z \in \mathbb{C} : z \text{ is a root of } \widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k \quad \text{with} \quad |||\widetilde{p} - p|||_2 \leq \epsilon |||p|||_2 \right\},$$

The pseudozero set $Z_\epsilon(p)$ can be characterized in terms of the level curves of a certain function [150, Proposition 2.1]:

$$Z_\epsilon(p) = \left\{ z \in \mathbb{C} : \quad \psi(z) = \frac{|p(z)|}{|||p|||_2 \|\Lambda(z)\|_2} \leq \epsilon \right\}, \tag{8.18}$$

where the vector $\Lambda(z)$ is defined in (8.3). This characterization of $Z_\epsilon(p)$ allows to determine it numerically.

Also recall, from Section 1.2.2.2, that the ϵ-pseudospectrum of M_σ, denoted by $\Lambda_\epsilon(M_\sigma)$, is the set

$$\Lambda_\epsilon(M_\sigma) := \{ z \in \mathbb{C} : z \text{ is an eigenvalue of } M_\sigma + E \text{ for some } E \text{ with } \|E\|_2 \leq \epsilon \|M_\sigma\|_2 \}.$$

The pseudospectrum set $\Lambda_\epsilon(M_\sigma)$ can be characterized in terms of the level curves of the norm of the resolvent [155, Theorem 2.2]:

$$\Lambda_\epsilon(M_\sigma) = \left\{ z : \|(zI - M_\sigma)^{-1}\|_2 \geq (\epsilon \|M_\sigma\|_2)^{-1} \right\}, \tag{8.19}$$

where by convention $\|(zI - M_\sigma)^{-1}\|_2$ takes the value ∞ in the spectrum of M_σ. Since the 2-norm of the resolvent matrix $(zI - M_\sigma)^{-1}$ is needed to compute $\Lambda_\epsilon(M_\sigma)$, we begin presenting in Theorem 8.21 an explicit expression of $(zI - M_\sigma)^{-1}$ for any Fiedler matrix.

Theorem 8.21. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let M_σ be the Fiedler matrix of $p(z)$ associated with the bijection σ, and let $z \in \mathbb{C}$ be a complex number that is not a root of $p(z)$. Then,*

$$(zI - M_\sigma)^{-1} = \frac{1}{p(z)} x_\sigma y_\sigma^T - A_\sigma = \frac{1}{p(z)} v_\sigma w_\sigma^T + B_\sigma, \tag{8.20}$$

where $x_\sigma, y_\sigma, v_\sigma, w_\sigma \in \mathbb{C}^n$ and $A_\sigma, B_\sigma \in \mathbb{C}^{n \times n}$ are defined in Theorem 5.3.

Proof. Using that, for any non singular matrix $A \in \mathbb{C}^{n \times n}$, $A^{-1} = \text{adj}(A)/\det(A)$ and that $\det(zI - M_\sigma) = p(z)$, Equation (8.20) is an immediate consequence of Theorem 5.3. □

8.5.1 Fast computation of pseudospectra of Fiedler matrices

Since $\|(zI - M_\sigma)^{-1}\|_2$ is equal to the inverse of the minimum singular value of $zI - M_\sigma$, one obvious way to determine $\Lambda_\epsilon(M_\sigma)$ numerically is to compute the minimum singular value of $zI - M_\sigma$, via the SVD, on a grid in the complex plane and, then, generate a contour plot from this data. The problem with this algorithm is that computing the whole SVD of a $n \times n$ matrix on a $m \times m$ grid requires $O(m^2 n^3)$ floating point operations, which is highly expensive. As it was commented in Section 1.2.2.2, different techniques have been introduced to make the computation of the norm of the resolvent matrix as efficient as possible. With these techniques the overall complexity can be reduced to $O(n^3 + n^2 m^2)$. Nevertheless, we will show that $\Lambda_\epsilon(M_\sigma)$ can be accurately estimated on a $m \times m$ grid in only $O(nm^2)$ flops. This result relies on Theorem 8.22, which shows that the ϵ-pseudospectrum of a Fiedler matrix is the region bounded by the ϵ-level curve of a certain function (denoted by $\phi_\sigma(z)$), that is easy and fast to compute, defined over the complex plane

when ϵ is sufficiently small. This result was proved in [150, Proposition 6.2] only when M_σ is one of the Frobenius companion matrices and it is extended here to any Fiedler matrix via the result in Theorem 8.21. We want to emphasize that Theorem 8.21 relies on Theorem 5.3. Hence, though the proof of Theorem 8.22 we provide here is rather short, it is based on strong technical results on Fiedler matrices.

Theorem 8.22. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let M_σ be the Fiedler companion matrix of $p(z)$ associated with σ, and let x_σ, y_σ, v_σ and w_σ be the vectors defined in Theorem 5.3. Then,*

$$\frac{\|(zI - M_\sigma)^{-1}\|_2}{\phi_\sigma(z)} = 1 + O\left(\frac{1}{\phi_\sigma(z)}\right) \quad as \quad \phi_\sigma(z) \longrightarrow \infty, \tag{8.21}$$

where

$$\phi_\sigma(z) = \begin{cases} \|x_\sigma\|_2 \|y_\sigma\|_2 / |p(z)| & \text{if } |z| \leq 1, \quad \text{and} \\ \|v_\sigma\|_2 \|w_\sigma\|_2 / |p(z)| & \text{if } |z| > 1. \end{cases} \tag{8.22}$$

Proof. From Theorem 8.21, we get

$$\|(zI - M_\sigma)^{-1}\|_2 = \left\| \frac{x_\sigma y_\sigma^T}{p(z)} - A_\sigma \right\|_2 = \left\| \frac{v_\sigma w_\sigma}{p(z)} + B_\sigma \right\|_2. \tag{8.23}$$

Recall from Lemma 5.8, that the entries of A_σ and B_σ are polynomials in z and z^{-1}, respectively. Therefore, there exist finite constants $0 \leq m_1, m_2 < \infty$, such that

$$\|A_\sigma\|_2 \leq m_1 \quad \text{if} \quad |z| \leq 1 \quad \text{and} \quad \|B_\sigma\|_2 \leq m_2 \quad \text{if} \quad |z| \geq 1. \tag{8.24}$$

Finally, to prove the result, we have to distinguish two cases: when $|z| \leq 1$ and when $|z| > 1$.

(a) If $|z| \leq 1$, from (8.24) we get

$$\left| \|(zI - M_\sigma)^{-1}\|_2 - \frac{\|x_\sigma\|_2 \|y_\sigma\|_2}{|p(z)|} \right| \leq \|A_\sigma\|_2 \leq m_1.$$

(b) If $|z| > 1$, from (8.24) we get

$$\left| \|(zI - M_\sigma)^{-1}\|_2 - \frac{\|v_\sigma\|_2 \|w_\sigma\|_2}{|p(z)|} \right| \leq \|B_\sigma\|_2 \leq m_2.$$

Therefore,

$$\left| \frac{\|(zI - M_\sigma)^{-1}\|_2}{\phi_\sigma(z)} - 1 \right| \leq \frac{\max\{m_1, m_2\}}{\phi_\sigma(z)},$$

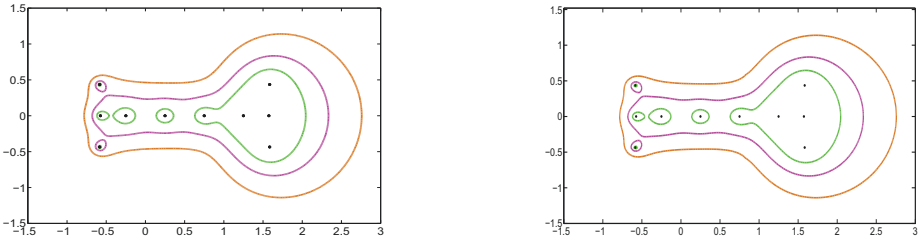
which implies (8.21). □

Remark 8.23. Recall from Theorems 8.2 and 5.3 that the entries of the vectors x_σ, y_σ, v_σ, and w_σ are either powers of z, or powers of z times a horner shift of $p(z)$ or $p^{\text{rev}}(z)$. This observation, together with the recurrence relation (2.6) to compute the Horner shifts of a polynomial and with the Horner's rule for evaluating polynomials, implies that $\phi(z)_\sigma$ can by computed in $O(n)$ flops.

Notice that in the neighborhood of a root of $p(z)$ we have $\phi_\sigma(z) \gg 1$. In this case, Theorem 8.22 shows that $\phi_\sigma(z)$ provides an accurate estimate of $\|(zI - M_\sigma)^{-1}\|_2$. Therefore, in the limit $\epsilon \to 0$, the pseudospectrum $\Lambda_\epsilon(M_\sigma)$ agrees with the region bounded by the $(\epsilon \|M_\sigma\|_2)^{-1}-$level curve of $\phi_\sigma(z)$. This result has practical applications since pseudospectra, as we said in Section 1.2.2.2 and in the paragraph before Theorem 8.22, are expensive to compute. As it is stated in Remark 8.23,

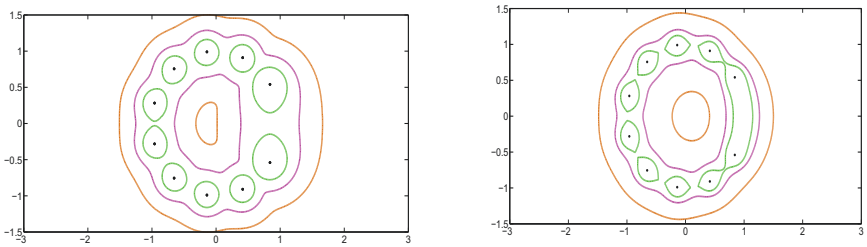
only $O(n)$ flops are needed to calculate $\phi_\sigma(z)$, as compared with $O(n^3)$ flops needed to calculate $\|(zI - M_\sigma)^{-1}\|_2$ by the SVD. Therefore, the function $\phi_\sigma(p)$ can be evaluated on a $m \times m$ grid in only $O(nm^2)$ flops.

We illustrate how the ϵ-pseudospectrum of a Fiedler matrix is accurately estimated by the $(\epsilon\|M_\sigma\|_2)^{-1}-$level curve of the function $\phi_\sigma(z)$ with three examples. In Figure 8.5.1, we plot, for $\epsilon = 10^{-2.5}, 10^{-3}, 10^{-3.5}$, in (a) the $\epsilon-$pseudospectra and in (b) the $(\epsilon\|M_\sigma\|_2)^{-1}-$level curves of the function $\phi_\sigma(z)$ of the Fiedler matrix $M_\sigma = C_2$ of the Bernoulli polynomial of degree 10: $z^{10} - 5z^9 + (15/2)z^8 - 7z^6 + 5z^4 - (3/2)z^2 + 5/66$. In Figure 8.5.2, we plot, for $\epsilon = 10^{-1.25}, 10^{-1}, 10^{-0.75}$, in (a) the $\epsilon-$pseudospectra and in (b) the $(\epsilon\|M_\sigma\|_2)^{-1}-$level curves of the function $\phi_\sigma(z)$, for the Fiedler matrix $M_\sigma = P_1$ defined in (2.7) of the polynomial $z^{10} + z^9 + \cdots + z + 1$. In Figure 8.5.3, we plot, for $\epsilon = 10^{-16}, 10^{-15}, 10^{-14}$, in (a) the $\epsilon-$pseudospectra and in (b) the $(\epsilon\|M_\sigma\|_2)^{-1}-$level curves of the function $\phi_\sigma(z)$, for the Fiedler matrix $M_\sigma = F$ defined in (2.8) of the monic polynomial with zeros in $1, 2, \ldots, 10$.



(a) (b)

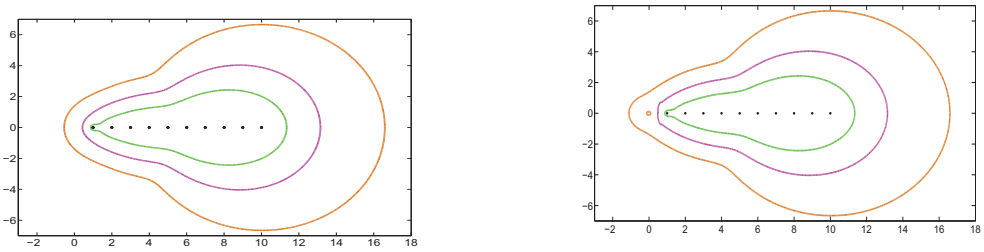
Figure 8.5.1: For the second Frobenius companion matrix $M_\sigma = C_2$ of the Bernoulli polynomial of degree 10: $z^{10} - 5z^9 + (15/2)z^8 - 7z^6 + 5z^4 - (3/2)z^2 + 5/66$, and for $\epsilon = 10^{-3.5}, 10^{-3}, 10^{-2.5}$, we plot in (a) the $\epsilon-$pseudospectra of C_2 and in (b) the $\epsilon-$ level curves of the function $\phi_\sigma(z)$ defined in (8.22), in green, magenta and brown, respectively.



(a) (b)

Figure 8.5.2: For the pentadiagonal Fiedler matrix $M_\sigma = P_1$ defined in (2.7) of the monic polynomial $z^{10} + z^9 + \cdots + z + 1$, and for $\epsilon = 10^{-1.25}, 10^{-1}, 10^{-0.75}$, we plot in (a) the $\epsilon-$pseudospectra of P_1 and in (b) the $\epsilon-$ level curves of the function $\phi_\sigma(z)$ defined in (8.22), in green, magenta and brown, respectively.

Notice that Figures 8.5.1-(a) and 8.5.1-(b), and 8.5.3-(a) and 8.5.3-(b) are almost indistinguishable. But also notice that, by contrast with Figures 8.5.1 and 8.5.3, there are some relevant differences between Figures 8.5.2-(a) and 8.5.2-(b). The main reason for these differences is that we are computing pseudospectra close to the region $|z| = 1$ where the 2-norm of the matrices A_σ and B_σ in Theorem 8.21 might not be negligible and, therefore, $\|(zI - M_\sigma)^{-1}\|_2 \approx \phi_\sigma(z)$ might not hold.



(a) (b)

Figure 8.5.3: For the Fiedler matrix $M_\sigma = F$ defined in (2.8) of the monic polynomial with zeros in $1, 2, \ldots, 10$, and for $\epsilon = 10^{-16}, 10^{-15}, 10^{-14}$, we plot in (a) the $\epsilon-$pseudospectra of F and in (b) the $\epsilon-$ level curves of the function $\phi_\sigma(z)$ defined in (8.22), in green, magenta and brown, respectively.

8.5.2 Asymptotic relations between pseudozero sets and pseudospectra of Fiedler matrices

Theorem 8.22 is also the key tool to prove the main results in this section, that is, Corollaries 8.24 and 8.25. These two corollaries give several asymptotic relations between the ϵ-pseudozero set of a monic polynomial $p(z)$ and the pseudospectrum of the Fiedler matrices of $p(z)$ in a neighborhood of a simple nonzero root λ of $p(z)$.

Corollary 8.24. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler companion matrix of $p(z)$ associated with σ. If λ is a simple nonzero root of $p(z)$, then*

$$\lim_{z \to \lambda} \frac{\|(zI - M_\sigma)^{-1}\|_2 \|M_\sigma\|_2}{1/\psi(z)} = \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)},$$

where $\psi(z) = |p(z)|/(\|\Lambda(z)\|_2 \|p\|_2)$ is the function in (8.18)*, and where $\Lambda(z)$ is defined in* (8.3)*.*

Proof. From Theorem 8.2 together with (8.6) and (8.22), we have

$$\lim_{z \to \lambda} \frac{\|M_\sigma\|_2}{|z|} \frac{|p(z)|\phi_\sigma(z)}{|p'(z)|} = \kappa(\lambda, M_\sigma). \tag{8.25}$$

Therefore,

$$\begin{aligned} \frac{\|(zI - M_\sigma)^{-1}\|_2 \|M_\sigma\|_2}{1/\psi(z)} &= \frac{\|(zI - M_\sigma)^{-1}\|_2}{\phi_\sigma(z)} \frac{|p'(z)| \cdot |z|}{\|\Lambda(z)\|_2 \|p\|_2} \frac{|p(z)|\phi_\sigma(z)\|M_\sigma\|_2}{|z| \cdot |p'(z)|} \\ &\xrightarrow[z \to \lambda]{} \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)}, \end{aligned}$$

where we have used (8.4), (8.25), and Theorem 8.22. ∎

In words, Corollary 8.24 says that, if M_σ is a Fiedler matrix of $p(z)$, then, in the limit $\epsilon \to 0$, the components of $\Lambda_\epsilon(M_\sigma)$ and $Z_{\epsilon'}(p)$ containing λ, where $\epsilon' = \epsilon\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, agree with each other (see (8.18) and (8.19)).

Corollary 8.24 together with Theorem 8.13 suggests that the ϵ-pseudospectrum of a Fiedler matrix of a monic polynomial $p(z)$ may be potentially much larger than the ϵ-pseudozero set of that polynomial when the maximum of the absolute values of the coefficients of $p(z)$ is large or close to zero. Nevertheless, Corollary 8.24 reveals that when $\max\{|a_{n-1}|, \ldots, |a_1|, |a_0|\}$ is moderate

and not close to zero, that is, it is of order $\Theta(1)$, the pseudozero sets of a monic polynomial and the pseudospectra of the associated Fiedler matrices will be quite close to each other for the same values of ϵ.

Corollary 8.25. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree $n \geq 2$, let $\sigma_1, \sigma_2 : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be two bijections, and let M_{σ_1} and M_{σ_2} be the Fiedler companion matrices of $p(z)$ associated with σ_1 and σ_2, respectively. Then, if λ is a nonzero simple root of $p(z)$,*

$$\lim_{z \to \lambda} \frac{\|(zI - M_{\sigma_1})^{-1}\|_2 \|M_{\sigma_1}\|_2}{\|(zI - M_{\sigma_2})^{-1}\|_2 \|M_{\sigma_2}\|_2} = \frac{\kappa(\lambda, M_{\sigma_1})}{\kappa(\lambda, M_{\sigma_2})}. \tag{8.26}$$

Proof. From (8.25), with $\sigma = \sigma_1, \sigma_2$, and Theorem 8.22, we have

$$\begin{aligned} \frac{\|(zI - M_{\sigma_1})^{-1}\|_2 \|M_{\sigma_1}\|_2}{\|(zI - M_{\sigma_2})^{-1}\|_2 \|M_{\sigma_2}\|_2} &= \frac{\phi_{\sigma_2}(z)\|(zI - M_{\sigma_1})^{-1}\|_2}{\phi_{\sigma_1}(z)\|(zI - M_{\sigma_2})^{-1}\|_2} \frac{\phi_{\sigma_1}(z)|p(z)|}{|p'(z)|} \frac{\|M_{\sigma_1}\|_2}{|z|} \frac{|p'(z)|}{\phi_{\sigma_2}(z)|p(z)|} \frac{|z|}{\|M_{\sigma_2}\|_2} \\ &\xrightarrow[z \longrightarrow \lambda]{} \frac{\kappa(\lambda, M_{\sigma_1})}{\kappa(\lambda, M_{\sigma_2})}. \end{aligned}$$

□

In words, Corollary 8.25 says that, if M_{σ_1} and M_{σ_2} are two different Fiedler matrices of $p(z)$, then, in the limit $\epsilon \to 0$, the components of $\Lambda_\epsilon(M_{\sigma_1})$ and $\Lambda_{\epsilon'}(M_{\sigma_2})$, where $\epsilon' = \epsilon\kappa(\lambda, M_{\sigma_1})/\kappa(\lambda, M_{\sigma_2})$, containing λ agree with each other.

Corollary 8.25 together with Theorem 8.17 suggest that the ϵ-pseudospectrum of a Fiedler matrix $M_\sigma \neq C_1, C_2$ may be potentially much larger (or much smaller) than the ϵ-pseudospectrum of the Frobenius companion matrices for polynomials that have large coefficients, since, in this case, the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ and $\kappa(\lambda, C)/\kappa(\lambda, M_\sigma)$ can be large (with $C = C_1, C_2$). Nevertheless, Corollary 8.25 reveals a sufficient condition for the pseudospectra of Fiedler matrices other than the Frobenius ones and the pseudospectra of the Frobenius companion matrices to be quite close to each other for the same values of ϵ. This condition is $\max\{|a_{n-1}|, \ldots, |a_1|, |a_0|\} = O(1)$.

8.6 Numerical experiments

In this section we provide numerical experiments that support our theoretical results. In particular, our goals are: (i) to show whether or not the bounds in Theorem 8.13 correctly predict the dependence on the coefficients of $p(z)$ of the largest ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ that may be obtained; (ii) to show whether or not the bounds in Theorem 8.17 correctly predict the dependence on the coefficients of $p(z)$ of the largest and smallest ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ that may be obtained, where C denotes one of the Frobenius companion matrices; (iii) to study the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ when the coefficients of $p(z)$ are bounded in absolute value by a moderate constant; and (iv) to investigate, from the point of view of condition numbers and pseudospectra, the effect of balancing Fiedler matrices. The reason to include the ratio $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ in the numerical experiments, although it is not studied in the previous sections, is that in [150] numerical experiments to study $\kappa(\lambda, C)/\text{cond}(\lambda, p)$ are provided, and, so, we extend that study to all Fiedler matrices.

Given a monic polynomial $p(z)$ of degree n, the second Frobenius companion matrix C_2 of $p(z)$, and a Fiedler matrix M_σ other than the Frobenius ones associated with $p(z)$, we are interested in the following quantities:

- $\max_\lambda \kappa(\lambda, C_2)/\kappa(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$,
- $\max_\lambda \kappa(\lambda, C_2)/\text{cond}(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$,

- $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, and
- $\min_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$,

where λ runs over all the nonzero simple roots of $p(z)$, and where the ratios of condition numbers $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\kappa(\lambda, C_2)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ are, respectively, the ratios in (8.8), (8.9) and (8.13), and where the ratio $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ is obtained from (8.4) and (8.6).

In the numerical experiments, we consider monic polynomials of degree 10 and the following Fiedler companion matrices associated with degree-10 polynomials:

(a) the second Frobenius companion matrix $C_2 = M_{\sigma_1}$ with $\text{PCIS}(\sigma_1) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$,

(b) the pentadiagonal Fiedler matrix $P_1 = M_{\sigma_2}$ with $\text{PCIS}(\sigma_2) = (1, 0, 1, 0, 1, 0, 1, 0, 1)$,

(c) the Fiedler matrix $F = M_{\sigma_3}$ with $\text{PCIS}(\sigma_3) = (0, 1, 1, 1, 1, 1, 1, 1, 1)$, and

(d) the Fiedler matrix M_{σ_4} with $\text{PCIS}(\sigma_4) = (0, 0, 1, 1, 1, 0, 1, 0, 0)$.

Recall that the matrices M_{σ_2} and M_{σ_3} are the Fiedler matrices considered in (2.7) and (2.8), respectively.

Given a monic polynomial $p(z)$ of degree 10, a Fiedler matrix M_σ associated with $p(z)$, and the second Frobenius companion matrix C_2, to compute the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ we proceed as follows. First, we compute the roots of $p(z)$ as the eigenvalues of the second Frobenius companion matrix C_2 with 64 digital digits of accuracy in MATLAB using the function `vpa` (variable precision arithmetic) followed by the command `eig`. Then, we compute $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ using (8.8)-(8.9) and (8.13), respectively, and we compute the ratio $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ using (8.4) and (8.6).

8.6.1 Numerical experiments that show the dependence of $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ on the coefficients of $p(z)$

In this subsection, we perform numerical experiments to determine whether or not the ratio of condition numbers $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ behave like Theorems 8.13 and 8.17 predict. In particular, we provide two sets of numerical experiments to study the dependence of $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ on $\|p\|_2$ and on the function $\rho(p)$ (defined in (8.10)), and a set of numerical experiments to study the dependence of $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ on $\|p\|_2$.

In the first set of numerical experiments we study the dependence of the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ on $\|p\|_2$. For this purpose, we consider a random sample of two hundred degree-10 monic polynomials $p(z) = z^{10} + \sum_{i=0}^{9} a_i z^i$ with coefficients of the form $a_i = c_i \times 10^{e_i}$, for $i = 0, 1, \ldots, 9$, where c_i and e_i are drawn from the uniform distributions on the intervals $[-1, 1]$ and $[0, 5]$, respectively. All the generated polynomials satisfy $\max\{|a_0|, |a_1|, \ldots, |a_9|\} \geq 1$, and, so, the function $\rho(p)$ satisfies $\rho(p) \leq \sqrt{2}$ for these polynomials. Then, Theorem 8.13 predicts that

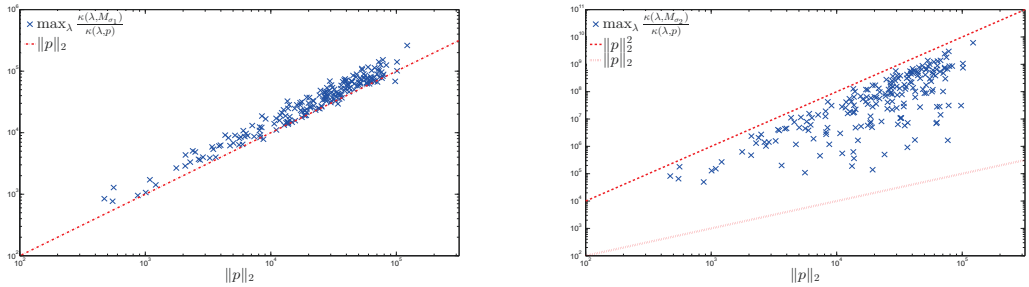
$$\frac{1}{\sqrt{2}} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq 10\sqrt{2}\|p\|_2 \tag{8.27}$$

if $M_\sigma = C_1, C_2$, and

$$\frac{1}{10} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq 100\sqrt{20}\|p\|_2^2 \tag{8.28}$$

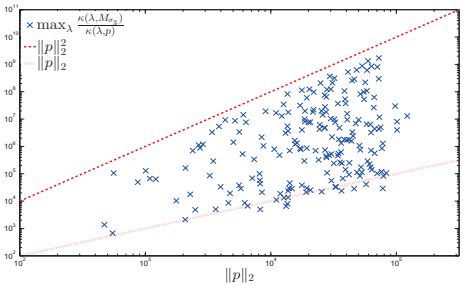
if $M_\sigma \neq C_1, C_2$.

In Figures 8.6.1-(a), 8.6.1-(b), 8.6.1-(c), and 8.6.1-(d) we plot for each of the 200 random polynomials the quantity $\max_\lambda \kappa(\lambda, M_{\sigma_i})/\kappa(\lambda, p)$, where λ runs over all nonzero simple roots of $p(z)$, for $i = 1, 2, 3, 4$, against the norm $\|p\|_2$. As may be seen in those figures, the largest ratios

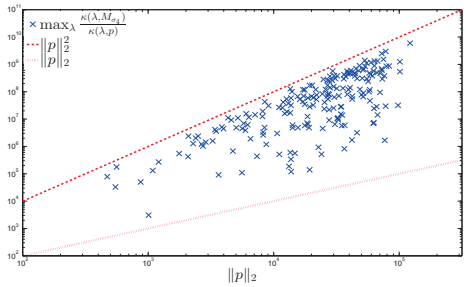


(a) M_{σ_1}

(b) M_{σ_2}



(c) M_{σ_3}



(d) M_{σ_4}

Figure 8.6.1: Maximum ratio $\kappa(\lambda, M_{\sigma_i})/\kappa(\lambda, p)$, for $i = 1, 2, 3, 4$, for each of the 200 random degree-10 monic polynomials with coefficients of the form $a_i = c_i \times 10^{e_i}$, for $i = 0, 1, \ldots, 9$, where c_i and e_i are drawn, respectively, from the uniform distributions on the intervals $[-1, 1]$ and $[0, 5]$.

obtained in these numerical experiments are bounded by a function that grows like $\|p\|_2$ in Figure 8.6.1-(a), and like $\|p\|_2^2$ in Figures 8.6.1-(b), 8.6.1-(c), and 8.6.1-(d). These results are consistent with the bounds in (8.27) and (8.28).

In the second set of numerical experiments we study the dependence of the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ on the function $\rho(p)$. For this purpose, we consider a random sample of two hundred degree-10 monic polynomials with coefficients close to zero. To generate those polynomials we proceed as follows. For $k = 1, 2, \ldots, 10$, we generate twenty degree-10 monic polynomials $p(z) = z^{10} + \sum_{i=0}^{9} a_i z^i$ with coefficients of the form $a_i = c_i \times 10^{e_i}$, for $i = 0, 1, \ldots, 9$, where c_i and e_i are drawn from the uniform distribution on the intervals $[-1, 1]$ and $[-k/2, -k/2 + 0.5]$, respectively. All the generated polynomials satisfy $\max\{|a_0|, |a_1|, \ldots, |a_9|\} \leq 1$, and, therefore, the their norms satisfy $\|p\|_2 \leq \sqrt{10}$. Then, Theorem 8.13 predicts that

$$\frac{1}{\sqrt{2}} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq 10^{3/2}\rho(p), \tag{8.29}$$

if $M_\sigma = C_1, C_2$, and

$$\frac{1}{10} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq 10^{7/2}\rho(p), \tag{8.30}$$

if $M_\sigma \neq C_1, C_2$.

In Figures 8.6.2-(a), 8.6.2-(b), 8.6.2-(c), and 8.6.2-(d) we plot for each of the 200 random polynomials the quantity $\max_\lambda \kappa(\lambda, M_{\sigma_i})/\kappa(\lambda, p)$, where λ runs over all nonzero simple roots of $p(z)$, for $i = 1, 2, 3, 4$, against the function $\rho(p)$. As may be seen in those figures, the ratios obtained in these numerical experiments grow like the function $\rho(p)$. These results are consistent with the bounds in (8.29) and (8.30). Also notice that the four plots in Figure 8.6.2 are almost

indistinguishable. This result is in accordance with Theorem 8.17, who predicts, from the point of view of conditioning, that for polynomials with moderate coefficients all Fiedler matrices behave like the Frobenius ones.

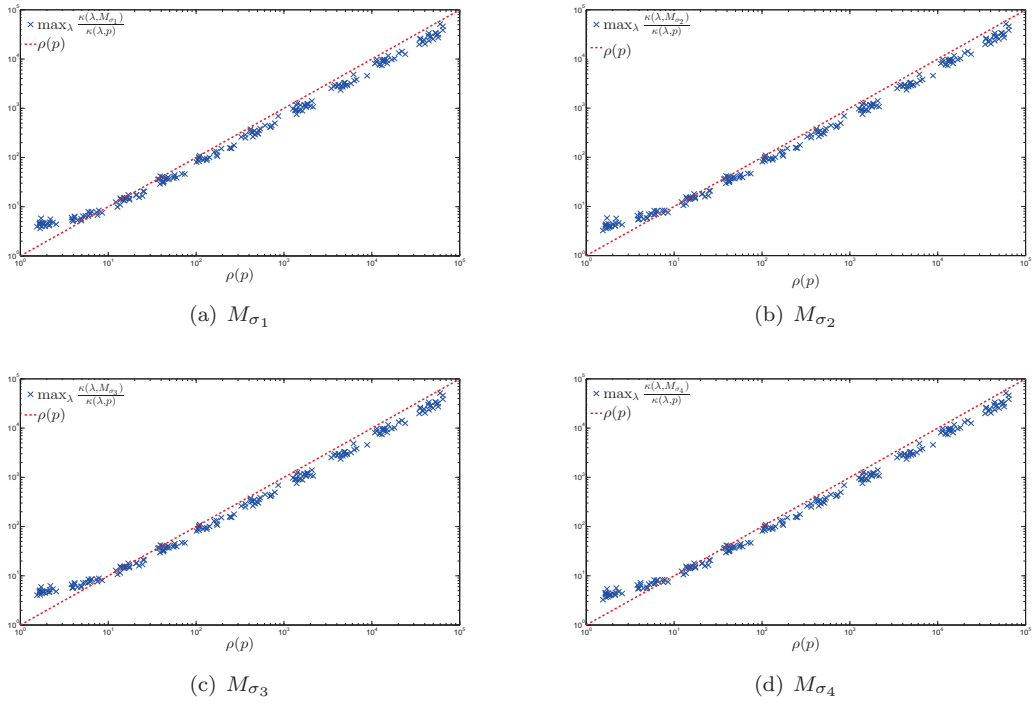


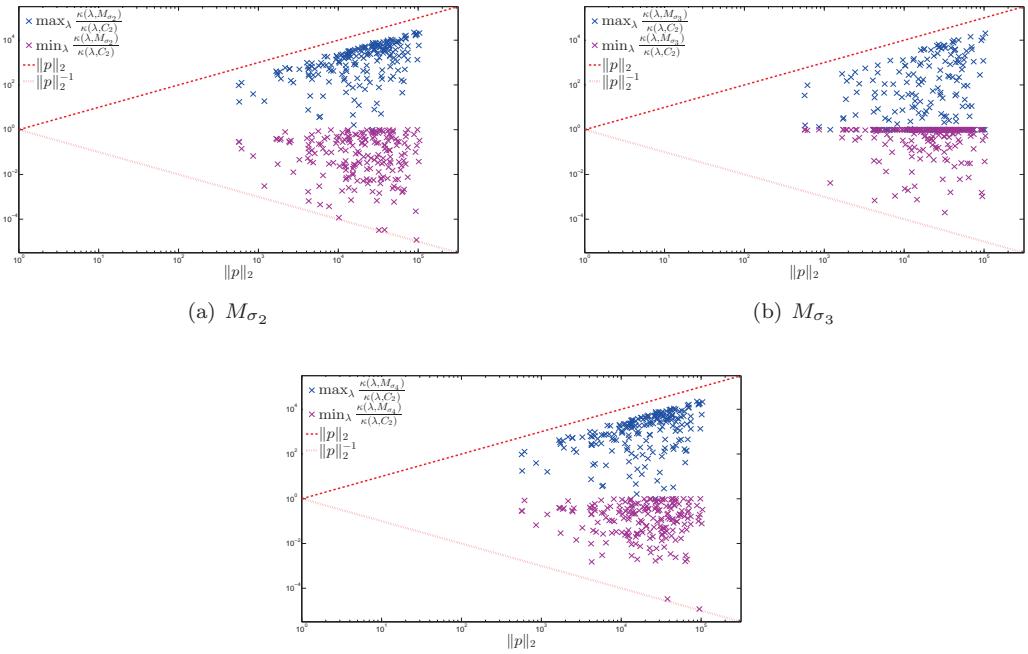
Figure 8.6.2: Maximum ratio $\kappa(\lambda, M_{\sigma_i})/\kappa(\lambda, p)$, for $i = 1, 2, 3, 4$, for each of the 20 random degree-10 monic polynomials of each of the 10 samples of random polynomials with coefficients of the form $a_i = c_i \times 10^{e_i}$, for $i = 0, 1, \ldots, 9$, where c_i and e_i are drawn, respectively, from the uniform distributions on the intervals $[-1, 1]$ and $[-k/2, -k/2 + 0.5]$, for $k = 1, 2, \ldots, 10$.

In the third set of numerical experiments we study the dependence of the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ on $\|p\|_2$. For this purpose, we consider again a random sample of two hundred degree-10 monic polynomials $p(z) = z^{10} + \sum_{i=0}^{9} a_i z^i$ with coefficients of the form $a_i = c_i \times 10^{e_i}$, for $i = 0, 1, \ldots, 9$, where c_i and e_i are drawn from the uniform distributions on the intervals $[-1, 1]$ and $[0, 5]$.

In Figures 8.6.3-(a), 8.6.3-(b), and 8.6.3-(c) we plot for each of the 200 random polynomials the quantities $\max_\lambda \kappa(\lambda, M_{\sigma_i})/\kappa(\lambda, C_2)$ and $\min_\lambda \kappa(\lambda, M_{\sigma_i})/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, for $i = 2, 3, 4$, against the norm $\|p\|_2$. As may be seen in those figures, the largest ratios obtained in these numerical experiments are upper bounded by a function that grows like $\|p\|_2$, and the smallest ratios are lower bounded by a function that decreases like $\|p\|_2^{-1}$. These results are consistent with the bounds in Theorem 8.17.

8.6.2 Numerical experiments with polynomials of moderate coefficients

In this subsection we study the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ when the coefficients of $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ are bounded in absolute value by a moderate number. In particular, we provide numerical evidence to show that: (i) the three ratios are moderate when all the coefficients or $p(z)$ are moderate and not close to zero; (ii) the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ are moderate when $\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}$ is moderate and not close to zero, but, in this situation, $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ may be large, and (iii) the ratio



(a) M_{σ_2} (b) M_{σ_3}

(c) M_{σ_4}

Figure 8.6.3: Maximum and minimum ratio $\kappa(\lambda, M_{\sigma_i})/\kappa(\lambda, C_2)$, in blue and purple, respectively, for $i = 1, 2, 3, 4$, for each of the 200 random degree-10 monic polynomials with coefficients of the form $a_i = c_i \times 10^{e_i}$, for $i = 0, 1, \ldots, 9$, where c_i and e_i are drawn, respectively, from the uniform distributions on the intervals $[-1, 1]$ and $[0, 5]$.

$\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ is moderate when $\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}$ is moderate, but, in this situation, the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ may be large.

In the first set of numerical experiments, we consider a random sample of 1000 degree-10 polynomials with coefficients drawn from the uniform distribution on the interval $[-10, 10]$, so that all the coefficients of every polynomial in the sample are moderate and not close to zero. In Table 8.6.1, we give the mean and the maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$.

As may be seen from the data in Table 8.6.1, the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, obtained for the polynomials in the first sample, are moderate, although, as may be seen from the data in Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$, the ratio $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ tends to be $O(10^3)$.

In the second set of numerical experiments, we consider a random sample of 1000 degree-10 polynomials with coefficients drawn from the uniform distribution on the interval $[-10, 10]$, but we set $a_0 = 10^{-10}$, so that all polynomials in this sample satisfy that $\max\{|a_0|, |a_1|, \ldots, |a_9|\}$ is moderate and not close to zero. The reason for setting $a_0 = 10^{-10}$ is to show that, in this situation, $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ is moderate but $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ may be large. In Table 8.6.2, we give the mean and the maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$.

As may be seen from the data in Table 8.6.2, the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, obtained for the polynomials in the second sample, are moderate, but, as may be seen from the data in Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$, the ratios $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ are large.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	1.5	2.0	1.9	2.1
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	1.8	2.6	2.4	2.6
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$	2.1	2.6	2.5	2.6
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$	4.1	3.1	3.0	3.2
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.0	0.5	0.4	0.6
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.0	0.8	0.8	0.8

Table 8.6.1: Mean and the maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for 1000 random degree-20 polynomials, with coefficients drawn from the uniform distribution on the interval $[-10, 10]$.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	1.5	1.9	1.4	1.9
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	1.7	2.4	1.7	2.4
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$	12.3	11.6	11.7	11.8
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$	12.6	12.0	12.2	12.2
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.0	0.5	0.0	0.5
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.0	0.7	0.0	0.8

Table 8.6.2: Mean and maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for 1000 random degree-10 polynomials, with coefficients drawn from the uniform distribution on the interval $[-10, 10]$ and setting $a_0 = 10^{-10}$.

In the third set of numerical experiments, we consider a random sample of 1000 degree-10 polynomials with coefficients of the form $a_i = c \cdot 10^e$ where c and e are drawn from the uniform distributions on the intervals $[-1, 1]$ and $[-10, -8]$, respectively, so that all polynomials in this sample satisfy that $\max\{|a_0|, |a_1|, \ldots, |a_9|\}$ is moderate but close to zero. The reason for this choice of random polynomials is to show that $\max\{|a_0|, |a_1|, \ldots, |a_9|\} = O(1)$ is not enough to guarantee that $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$ are moderate. In Table 8.6.3, we give the mean and the maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$.

As may be seen in Table 8.6.3, the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, obtained for the polynomials in the third sample of random polynomials, are moderate, but, as may be seen from the data in Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$ and Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\mathrm{cond}(\lambda, p)$ are large.

8.6.3 Numerical experiments balancing Fiedler matrices

Given a polynomial $p(z)$ and an associated Fiedler companion matrix M_σ, it would be desirable to find a similarity transformation that makes the eigenvalue problem no worse conditioned that the polynomial root-finding problem. If D is a nonsingular diagonal matrix, then the condition

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	8.3	8.3	8.3	8.3
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	9.2	9.2	9.2	9.2
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$	9.4	9.4	9.4	9.4
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$	12.9	12.9	12.9	12.9
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.0	0.0	0.0	0.0
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.0	0.0	0.0	0.0

Table 8.6.3: Mean and maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for 1000 random degree-10 polynomials, with coefficients of the form $a_i = c \cdot 10^e$ where c and e are drawn from the uniform distributions on the intervals $[-1, 1]$ and $[-10, -8]$, respectively.

number $\kappa(\lambda, DM_\sigma D^{-1})$ of a nonzero simple eigenvalue λ of M_σ is given by

$$\kappa(\lambda, DM_\sigma D^{-1}) = \frac{\|DM_\sigma D^{-1}\|_2}{|\lambda|} \frac{\|Dr_\sigma\|_2 \|D^{-1} l_\sigma\|_2}{|p'(\lambda)|}, \tag{8.31}$$

where the vectors r_σ and l_σ are defined in Theorem 8.2. In this subsection we perform numerical experiments to study, from the point of view of condition numbers and pseudospectra, the effect of balancing Fiedler matrices (see Section 1.2.3).

In the first set of numerical experiments, we consider a random sample of 1000 degree-10 monic polynomials $p(z)$ as in (1.1) with coefficients of the form $a_k = b_1 \times 10^{e_1} + ib_2 \times 10^{e_2}$, where, for $i = 1, 2$, b_i and e_i are drawn from the uniform distributions on the intervals $[-1, 1]$ and $[-10, 10]$, respectively. Our goal is to study the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$, $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ when the Fiedler matrices are not or are balanced. To compute the diagonal matrix D that balance a Fiedler matrix M_σ we use the command `balance` in MATLAB.

In Table 8.6.4, we give the mean and the maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, and the mean and the minimum of the decimal logarithms (Log-Mean and Log-Minimum, respectively) of $\min_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, for the Fiedler matrices $M_\sigma = M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, with and without balancing them.

Several observations may be drawn from the data in Tables 8.6.4-(a) and 8.6.4-(b). First note, from the data in Table 8.6.4-(a), that if the Fiedler matrices are not balanced, the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ may be large or small, as it is predicted in Theorem 8.17. Also note that the largest and smallest of these ratios are consistent with the bounds in (8.14) and (8.15). Second, note, from the data in Table 8.6.4-(b), that the process of balancing the Fiedler matrices makes the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$ moderate, so that, from the point of view of conditioning, balanced Fiedler Matrices can be used with the same reliability as balanced Frobenius companion matrices even with polynomials with large norms.

In Table 8.6.5, we give the mean and the maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$, where λ runs over all nonzero simple roots of $p(z)$, for the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, with and without balancing them.

Again, several conclusions may be drawn from the data in Tables 8.6.5-(a) and 8.6.5-(b). First note, from the point of view of conditioning, that if the Fiedler matrices are not balanced, the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ may be large, as it is predicted in Theorem 8.13. Note also that the largest of these ratios is consistent with the upper bound in (8.11) for the Frobenius companion matrix, and with the upper bound in (8.12) for Fiedler matrices other than the Frobenius ones. Second, note, comparing the data in Table 8.6.5-(a) with the data in Table 8.6.5-(b),

(a) The Fiedler matrices are not balanced.

	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	6.3	3.7	6.1
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	9.4	9.4	9.4
Log-Mean $\min_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	-4.8	-2.7	-5.4
Log-Minimum $\min_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	-9.5	-8.4	-9.3

(b) The Fiedler matrices are balanced.

	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.1	0.1	0.1
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	0.9	1.2	1.3
Log-Mean $\min_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	-0.7	-0.1	-0.6
Log-Minimum $\min_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$	-2.6	-0.9	-2.3

Table 8.6.4: Mean and maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, and mean and minimum of the decimal logarithms (Log-Mean and Log-Minimum, respectively) of $\min_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, C_2)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for a sample of 1000 random degree-10 monic polynomials, with coefficients of the form $b_1 \times 10^{e_1} + ib_2 \times 10^{e_2}$, where, for $i = 1, 2$, b_i and e_i are drawn from the uniform distributions on the intervals $[-1, 1]$ and $[-10, 10]$, respectively, for the Fiedler matrices $M_\sigma = M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing and with balancing.

that the process of balancing Fiedler matrices may reduce considerably the ratios $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$. In particular, from the data in Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ in Table 8.6.5-(b), we can see that, *usually*, balancing makes the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ moderate, although, from the data in Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ in Table 8.6.5-(b), we can see that balancing is not always enough to guarantee a moderate ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$. Finally, from the data in Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ and Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ in Table 8.6.5-(b), we see that balancing *in most cases* is not enough to guarantee that the ratio $\kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$ is moderate.

8.6.4 Numerical experiments to study pseudospectra of Fiedler matrices

In this section we perform some numerical experiments to illustrate the theoretical results in Section 8.5 concerning pseudospectra of Fiedler matrices, and to study the effect of balancing Fiedler matrices from the point of view of pseudospectra.

In the first numerical experiment, we illustrate Corollary 8.24, that is, in a neighborhood of a nonzero simple root λ of a monic polynomial $p(z)$, the pseudospectrum $\Lambda_{\epsilon'}(M_\sigma)$ and the pseudozero set $Z_\epsilon(p)$ containing λ, where $\epsilon' = \epsilon\kappa(\lambda, p)/\kappa(\lambda, M_\sigma)$, agree with each other.

In Figure 8.6.4 we plot, for $\epsilon = 10^{-10.75}, 10^{-10.5}, 10^{-10.25}$, the ϵ-pseudozero sets $Z_\epsilon(p)$, and, for $i = 1, 2, 3$, the ϵ_i-pseudospectra $\Lambda(M_{\sigma_i})$, where $p(z)$ is the monic polynomial $p(z) = \prod_{j=1}^{10}(z - j)$, and, for $i = 1, 2, 3$, $\epsilon_i = \epsilon\kappa(4, p)/\kappa(4, M_{\sigma_i})$. The ratios $\kappa(4, M_{\sigma_1})/\kappa(4, p) = 3.97 \cdot 10^6$, $\kappa(4, M_{\sigma_2})/\kappa(4, p) = 1.18 \cdot 10^9$ and $\kappa(4, M_{\sigma_3})/\kappa(4, p) = 5.13 \cdot 10^7$are computed using (8.8) and (8.9) in MATLAB. As can be seen in those figures, the pseudozero sets and the pseduospectra of the three Fiedler matrices are almost identical.

In the second numerical experiment, we present a graphical comparison between the pseudozero sets of a monic polynomial $p(z)$ with a large norm $\|p\|_2$, and the pseudospectra of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}$ associated with $p(z)$, to show that when $\|p\|_2$ is large there may be relevant differences between pseudozero sets and pseudospectra of Fiedler matrices.

In Figures 8.6.5-(a), 8.6.5-(b), 8.6.5-(c), and 8.6.5-(d) we plot, for $\epsilon = 10^{-11}, 10^{-12}, 10^{-13}$, the

(a) The Fiedler matrices are not balanced.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	8.6	14.6	11.8	14.4
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	10.4	19.4	19.4	19.3
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$	15.2	17.2	17.1	17.4
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$	28.6	27.9	27.9	28.5

(b) The Fiedler matrices are balanced.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	2.7	2.4	2.8	2.5
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$	7.9	7.5	8.2	7.7
Log-Mean $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$	7.4	6.9	7.5	7.0
Log-Maximum $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$	21.5	21.6	21.5	21.6

Table 8.6.5: Mean and maximum of the decimal logarithms (Log-Mean and Log-Maximum, respectively) of $\max_\lambda \kappa(\lambda, M_\sigma)/\kappa(\lambda, p)$ and $\max_\lambda \kappa(\lambda, M_\sigma)/\text{cond}(\lambda, p)$, where λ runs over all nonzero simple roots of $p(z)$, obtained for a sample of 1000 random degree-10 monic polynomials, with coefficients of the form $b_1 \times 10^{e_1} + ib_2 \times 10^{e_2}$, where, for $i = 1, 2$, b_i and e_i are drawn from the uniform distributions on the intervals $[-1, 1]$ and $[-10, 10]$, respectively, for the Fiedler matrices $M_\sigma = M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing and with balancing.

ϵ-pseudozero sets $Z_\epsilon(p)$ of the monic polynomial $p(z) = \prod_{j=1}^{10}(z - j)$, and the ϵ-pseudospectra $\Lambda_\epsilon(M_{\sigma_1}), \Lambda_\epsilon(M_{\sigma_2}), \Lambda_\epsilon(M_{\sigma_3})$ of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}$ associated with $p(z)$. As may be seen in those figures, the pseudozero sets and the pseudospectra of the Fiedler matrices are very different. Notice also that there are relevant differences between the pseudospectra $\Lambda_\epsilon(M_{\sigma_1})$, $\Lambda_\epsilon(M_{\sigma_2})$ and $\Lambda_\epsilon(M_{\sigma_3})$, although, for a fixed ϵ, the three pseudospectra have, approximately, the same size.

In the final set of numerical experiments, as in [53] and [150], we explore the following degree-20 monic polynomials:

($p1$) the Wilkinson polynomial: $p(z) = \prod_{k=1}^{20}(z - k)$,

($p2$) the monic polynomial with zeros: $-2, -1.8, -1.6, \ldots, 1.6, 1.8$,

($p3$) $p(z) = (20!) \sum_{k=0}^{20} z^k/k!$,

($p4$) the Bernoulli polynomial of degree 20:

$$\begin{aligned} p(z) =& z^{20} - 10z^{19} + \frac{95}{3}z^{18} - \frac{323}{2}z^{16} + \frac{6460}{7}z^{14} - 4199z^{12} + \frac{41990}{3}z^{10} - \frac{223193}{7}z^8 + 45220z^6 \\ & - \frac{68723}{2}z^4 + \frac{219335}{21}z^2 - \frac{174611}{330}, \end{aligned}$$

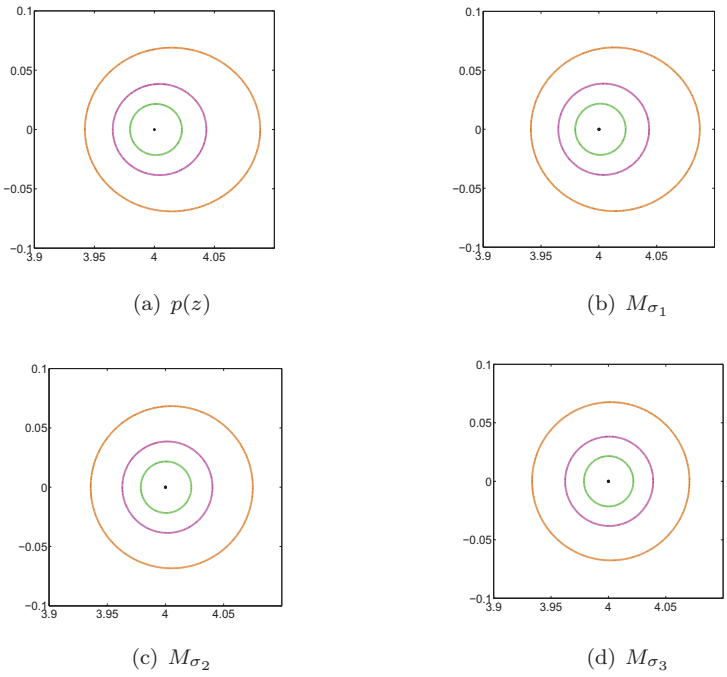
($p5$) $p(z) = \sum_{k=0}^{20} z^k$,

($p6$) the monic polynomial with zeros $2^{-10}, 2^{-9}, \ldots, 2^8, 2^9$,

($p7$) the Chebyshev polynomial of degree 20,

($p8$) the monic polynomial with zeros equally spaced on a sine curve, that is,

$$p(z) = \prod_{k=-10}^{9} \left(z - \frac{2\pi}{19}(k + 0.5) - i \cdot \sin \frac{2\pi}{19}(k + 0.5) \right).$$



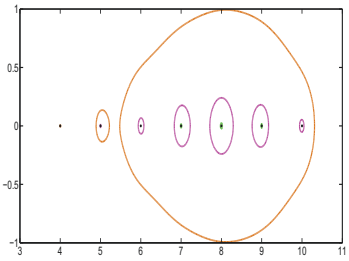
(a) $p(z)$ (b) M_{σ_1}

(c) M_{σ_2} (d) M_{σ_3}

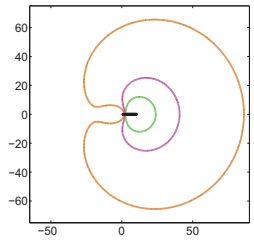
Figure 8.6.4: For $p(z) = \prod_{j=1}^{10}(z-j)$ and for $\epsilon = 10^{-10.75}, 10^{-10.5}, 10^{-10.25}$ we plot, in green, magenta and brown, respectively, the ϵ-pseudozero set $Z_\epsilon(p)$ and, for $i = 1, 2, 3$, the ϵ_i-pseudospectra $\Lambda(M_{\sigma_i})$, where $\epsilon_i = \epsilon\kappa(4, p)/\kappa(4, M_{\sigma_i})$, $M_{\sigma_1} = C_2$ is the second Frobenius companion matrix, $M_{\sigma_2} = P_1$ is the Fiedler matrix defined in (2.7), and $M_{\sigma_3} = F$ is the Fiedler matrix defined in (2.8).

As in [53], we first compute the coefficients exactly or with high precision using Mathematica. We then read these coefficients into MATLAB and take the rounded coefficients stored in MATLAB as our official test cases. Also, we consider again the Fiedler matrices $M_{\sigma_1} = C_2$ and $M_{\sigma_2} = P_2$, defined in (2.7), but, this time, associated with degree-20 polynomials.

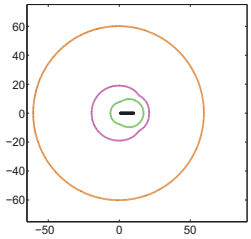
For each of the eight polynomials $p1$–$p8$ we present in Figures 8.6.6–8.6.13 a graphical comparison between the pseudozero sets $Z_\epsilon(p)$, the coefficientwise pseudozero sets $\text{Pseudo}_\epsilon(p)$ (see (1.26)) and the pseudospectra $\Lambda_\epsilon(\widetilde{M}_{\sigma_i})$, for $i = 1, 2$, where $\widetilde{M}_{\sigma_i}$ denotes the Fiedler matrix M_{σ_i} after being balanced. The goal of these comparisons is to show that, for *some* polynomials, balancing tends to achieve a reasonably close agreement between the coefficientwise pseudozero set $\text{Pseudo}_\epsilon(p)$ and the pseudospectra of balanced Fiedler matrices, at least compared with the size of the pseudozero set $Z_\epsilon(p)$. These figures also show, as can be seen by comparing Figures 8.6.6-(a)–8.6.13-(a) with Figures 8.6.6-(b)–8.6.13-(b), that pseudospectra of balanced Fiedler matrices other than the Frobenius ones may be larger that the pseudospectra of balanced Frobenius companion matrices, for the same values of ϵ. In other words, eigenvalues of balanced Fiedler matrices other than the Frobenius ones may be more sensitive to finite perturbations than eigenvalues of balanced Frobenius matrices. This result is in contrast with the numerical experiments in Section 8.6.3, which show that, under infinitesimal perturbations, the sensitivity of the eigenvalues of a balanced Fiedler matrix and the sensitivity of the eigenvalues of a balanced Frobenius matrix are approximately the same.



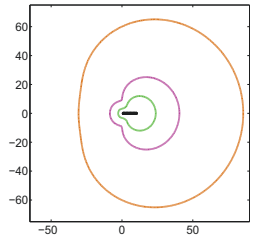
(a) $p(z)$



(b) M_{σ_1}

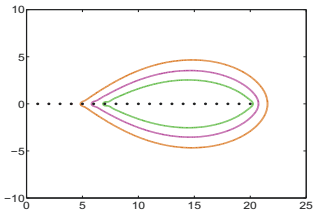


(c) M_{σ_2}

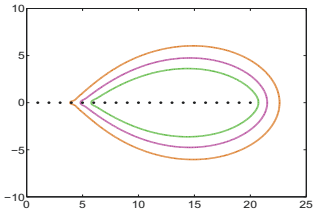


(d) M_{σ_3}

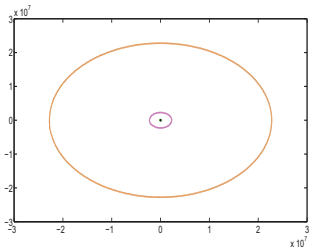
Figure 8.6.5: For $p(z) = \prod_{j=1}^{10}(z-j)$ and for $\epsilon = 10^{-11}, 10^{-12}, 10^{-13}$ we plot, in green, magenta and brown, respectively, the ϵ-pseudozero set $Z_\epsilon(p)$ and, for $i = 1, 2, 3$, the ϵ-pseudospectra $\Lambda(M_{\sigma_i})$.



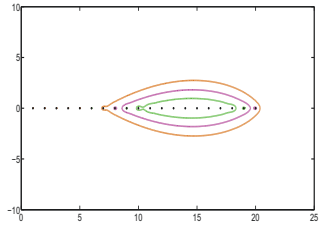
(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$



(b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

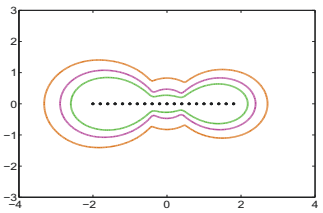


(c) $Z_\epsilon(p1)$

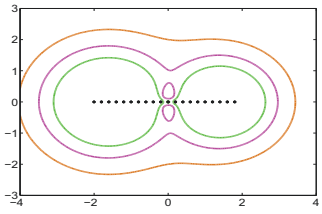


(d) Pseudo$_\epsilon(p1)$

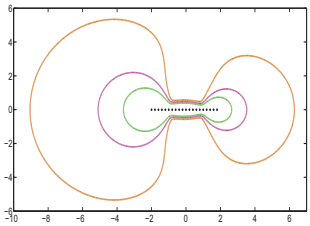
Figure 8.6.6: For $\epsilon = 10^{-14}, 10^{-13}, 10^{-12}$, we plot $Z_\epsilon(p1)$, Pseudo$_\epsilon(p1)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$, in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p1$.



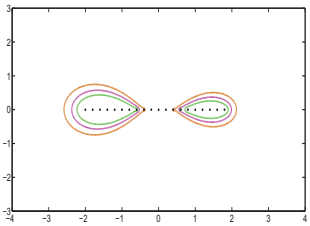
(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$



(b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

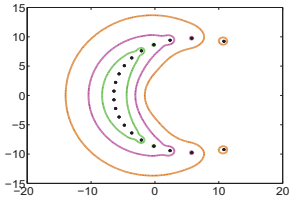


(c) $Z_\epsilon(p2)$

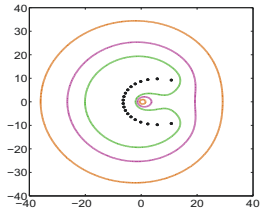


(d) $\text{Pseudo}_\epsilon(p2)$

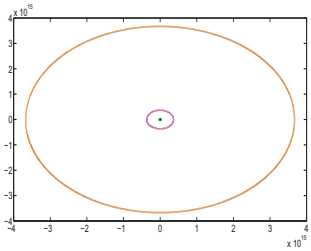
Figure 8.6.7: For $\epsilon = 10^{-3}, 10^{-2.5}, 10^{-2}$, we plot $Z_\epsilon(p2)$, $\text{Pseudo}_\epsilon(p2)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$, in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p2$.



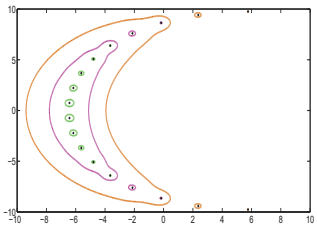
(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$



(b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

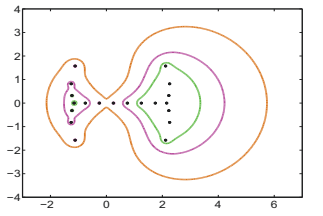


(c) $Z_\epsilon(p3)$

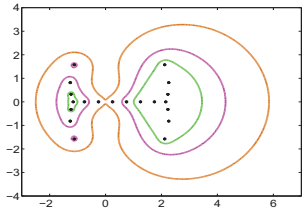


(d) $\text{Pseudo}_\epsilon(p3)$

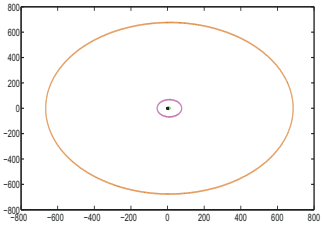
Figure 8.6.8: For $\epsilon = 10^{-5}, 10^{-4}, 10^{-3}$, we plot $Z_\epsilon(p3)$, $\text{Pseudo}_\epsilon(p3)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$, in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p3$.



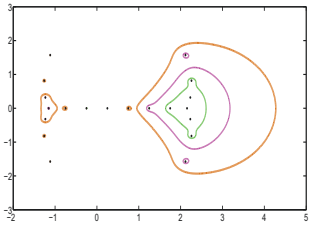
(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$



(b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

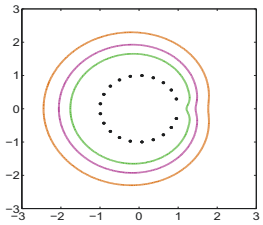


(c) $Z_\epsilon(p4)$

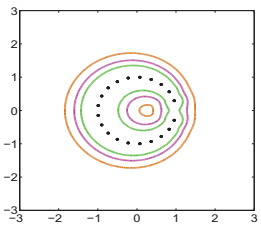


(d) Pseudo$_\epsilon(p4)$

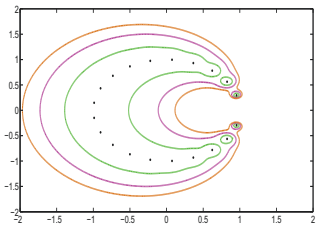
Figure 8.6.9: For $\epsilon = 10^{-4}, 10^{-3}, 10^{-2}$, we plot $Z_\epsilon(p4)$, Pseudo$_\epsilon(p4)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$, in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p4$.



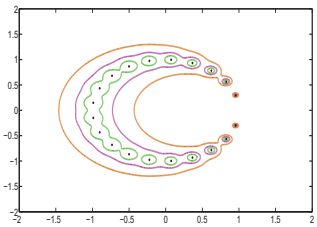
(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$



(b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

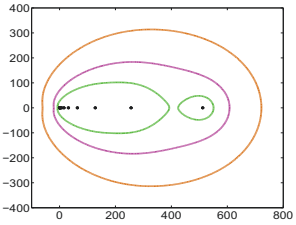


(c) $Z_\epsilon(p5)$

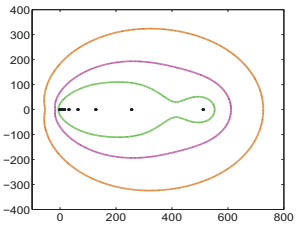


(d) Pseudo$_\epsilon(p5)$

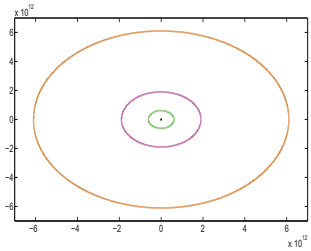
Figure 8.6.10: For $\epsilon = 10^{-1.2}, 10^{-1}, 10^{-0.8}$, we plot $Z_\epsilon(p5)$, Pseudo$_\epsilon(p5)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$,in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p5$.



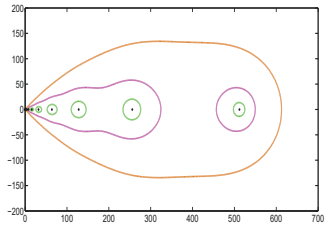
(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$



(b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

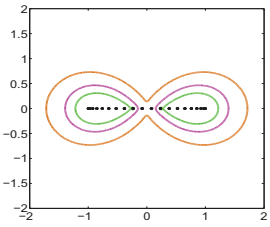


(c) $Z_\epsilon(p6)$

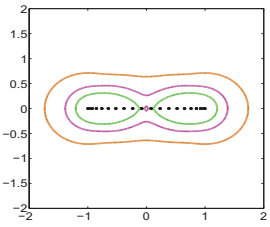


(d) Pseudo$_\epsilon(p6)$

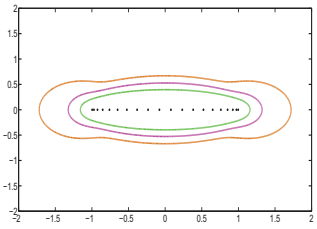
Figure 8.6.11: For $\epsilon = 10^{-2.5}, 10^{-2}, 10^{-1.5}$, we plot $Z_\epsilon(p6)$, Pseudo$_\epsilon(p6)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$, in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p6$.



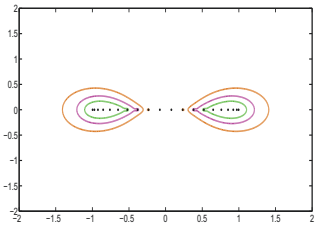
(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$



(b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

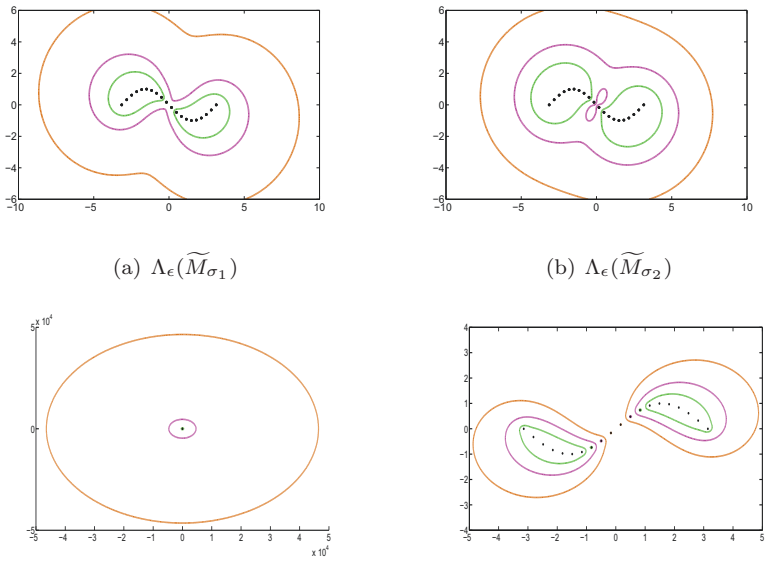


(c) $Z_\epsilon(p7)$



(d) Pseudo$_\epsilon(p7)$

Figure 8.6.12: For $\epsilon = 10^{-4}, 10^{-3}, 10^{-2}$, we plot $Z_\epsilon(p7)$, Pseudo$_\epsilon(p7)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$, in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p7$.



(a) $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ (b) $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$

(c) $Z_\epsilon(p8)$ (d) $\text{Pseudo}_\epsilon(p8)$

Figure 8.6.13: For $\epsilon = 10^{-3}, 10^{-2}, 10^{-1}$, we plot $Z_\epsilon(p8)$, $\text{Pseudo}_\epsilon(p8)$, $\Lambda_\epsilon(\widetilde{M}_{\sigma_1})$ and $\Lambda_\epsilon(\widetilde{M}_{\sigma_2})$, in green, magenta and brown, respectively, where $\widetilde{M}_{\sigma_1}$ and $\widetilde{M}_{\sigma_2}$ denote the balanced Fiedler matrices coming from M_{σ_1} and M_{σ_2}, respectively, of the polynomial $p8$.

Chapter 9

Backward stability of polynomial root-finding from Fiedler matrices

In this chapter, we investigate the backward error of the computed roots of a monic polynomial $p(z)$ when they are computed as the eigenvalues of a Fiedler matrix M_σ using a backward stable eigenvalue algorithm. The definition of the normwise backward error of the whole ensemble of computed roots, $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$, of $p(z)$ via a certain algorithm is

$$\eta(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n) := \frac{\|\widetilde{p} - p\|}{\|p\|},$$

for some polynomial norm, and where $\widetilde{p}(z) = \prod_{i=1}^{n}(z - \widetilde{\lambda}_i)$. Note that this notion of backward error coincides with the relative distance between the original polynomial $p(z)$ and the monic polynomial $\widetilde{p}(z)$ whose roots are $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$.

We also study whether or not computing the roots of $p(z)$ using a Fiedler matrix of $p(z)$ and a backward stable eigenvalue algorithm is backward stable from the point of view of the polynomials, that is, whether or not

$$\eta(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n) = O(u)$$

holds, where u is the machine epsilon. This work is motivated by [53, 101, 102, 160], which address related issues for the Frobenius companion matrices associated with scalar and matrix polynomials.

Throughout this chapter, if $A \in \mathbb{C}^{n\times n}$ is a matrix, then $\|A\|_\infty$ denotes the usual matrix ∞-norm (see [79, p. 108]). In particular, for a column vector $v = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^T \in \mathbb{C}^{n\times 1}$, we have $\|v\|_\infty = \max\{|v_1|, \ldots, |v_n|\}$, and for a row vector $v = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \in \mathbb{C}^{1\times n}$, we have $\|u\|_\infty = |u_1| + \cdots + |u_n|$. Similarly, for a polynomial $p(z) = \sum_{k=0}^{n} a_k z^k$ (not necessarily monic), $\|p\|_\infty$ is the norm on the vector space of scalar polynomials of degree less than or equal to n defined as

$$\|p\|_\infty := \max\{|a_n|, |a_{n-1}|, \ldots, |a_1|, |a_0|\}.$$

Notice that, since we deal in this dissertation with chapter polynomials, $a_n = 1$ and we always have $\|p\|_\infty \geq 1$.

9.1 Backward error of the computed roots using Fiedler matrices

Given a monic polynomial $p(z)$, if $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$ are the computed roots of $p(z)$ using a Fiedler matrix and a backward stable eigenvalue algorithm, the goal of this section is to bound the normwise

backward error

$$\eta_\infty(\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_n) := \frac{\|\widetilde{p}-p\|_\infty}{\|p\|_\infty},$$

where $\widetilde{p}(z)=\prod_{i=1}^n(z-\widetilde{\lambda}_i)$. Following the discussion in Section 1.2.1, we have that the computed roots $\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_n$ are the exact eigenvalues of a certain perturbation of M_σ, say $M_\sigma+E$, with E satisfying

$$\|E\|_\infty = O(u)\|M_\sigma\|_\infty,$$

and where u is the machine epsilon. In other words, we have $p(z)=\det(zI-M_\sigma)$ and $\widetilde{p}(z)=\det(zI-M_\sigma-E)$. Hence, the difference between $p(z)$ and $\widetilde{p}(z)$ can be measured from the variation of the coefficients of the characteristic polynomial of M_σ under a small perturbation of M_σ.

Thus, if we consider the kth coefficient of the characteristic polynomial of a matrix $X=(x_{ij})\in\mathbb{C}^{n\times n}$ as a function of the entries of X, that is, $a_k(X):\mathbb{C}^{n^2}\to\mathbb{C}$, for $k=0,1,\ldots,n-1$, then, to first order in E, we have (see (1.8))

$$|a_k(M_\sigma+E)-a_k(M_\sigma)| = |\nabla a_k(M_\sigma)\cdot \text{vec }(E)|, \quad \text{for } k=0,1,\ldots,n-1, \tag{9.1}$$

where $\text{vec}(E)$ is the column vector which vetorizes E (see (1.9)) and, for $k=0,1,\ldots,n-1$, $\nabla a_k(M_\sigma)$ is the gradient of the kth coefficient of the characteristic polynomial of a matrix evaluated at M_σ (see (1.10)).

The following well-know result (known as Jacobi's formula, see [15]) provides us a description of the gradient of the determinant.

Lemma 9.1. *Let $A\in\mathbb{C}^{n\times n}$ and consider a small perturbation $A+E$ of A, with $E\in\mathbb{C}^{n\times n}$. Then, the function*

$$\begin{array}{rccc}\det: & \mathbb{C}^{n\times n} & \longrightarrow & \mathbb{C}\\ & X & \longmapsto & \det(X),\end{array}$$

is analytic in a neighborhood of A, and

$$\det(A+E)=\det(A)+\text{tr}(\text{adj}(A)E)+O(\|E\|^2),$$

where $\|\cdot\|$ is any norm in $\mathbb{C}^{n\times n}$, $\text{adj}(A)$ denotes the adjugate matrix of A (see Definition 5.1)*, and $\text{tr}(B)$ denotes the trace of B.*

As an immediate consequence of Lemma 9.1, applied to $p(z)=\det(zI-A)$, we get Proposition 9.2, which gives a description of the gradient of the coefficients of the characteristic polynomial of any matrix A and, as a consequence, an expression for the variation of the characteristic polynomial under small perturbations, up to first order. Observe that Lemma 9.1 and Proposition 9.2 are valid for general matrices A and not only for M_σ.

Proposition 9.2. *Let $A\in\mathbb{C}^{n\times n}$ and $z\in\mathbb{C}$. Let us write the adjugate matrix of $zI-A$ as*

$$\text{adj}(zI-A)=\sum_{k=0}^{n-1} z^k P_{k+1}, \tag{9.2}$$

with $P_{k+1}\in\mathbb{C}^{n\times n}$, for $k=0,1,\ldots,n-1$. Let $a_k(X):\mathbb{C}^{n^2}\to\mathbb{C}$ be the kth coefficient of the characteristic polynomial of a matrix $X=(x_{ij})\in\mathbb{C}^{n\times n}$, and let $\nabla a_k(A)$ be the gradient of the function $a_k(X)$ evaluated at A. Then, for $k=0,1,\ldots,n-1$,

$$\nabla a_k(A) = -\left[\text{vec }(P_{k+1}^T)\right]^T.$$

As a consequence, if $A+E$ is a small perturbation of A, with $E \in \mathbb{C}^{n\times n}$, then

$$\begin{aligned}\det(zI-(A+E)) - \det(zI-A) &= -\sum_{k=0}^{n-1} z^k \left[\mathrm{vec}(P_{k+1}^T)\right]^T \cdot \mathrm{vec}(E) + O(\|E\|^2) \\ &= -\sum_{k=0}^{n-1} z^k \,\mathrm{tr}(P_{k+1}E) + O(\|E\|^2),\end{aligned}$$

where $\|\cdot\|$ is any norm in $\mathbb{C}^{n\times n}$.

Proof. From Lemma 9.1 and (9.2), we have

$$\begin{aligned}\det(zI-(A+E)) &= \det(zI-A) - \mathrm{tr}(\mathrm{adj}(zI-A)E) + O(\|E\|^2) \\ &= \det(zI-A) - \textstyle\sum_{k=0}^{n-1} z^k \mathrm{tr}(P_{k+1}E) + O(\|E\|^2) \\ &= \det(zI-A) - \textstyle\sum_{k=0}^{n-1} z^k \left[\mathrm{vec}(P_{k+1}^T)\right]^T \cdot \mathrm{vec}(E) + O(\|E\|^2),\end{aligned}$$

and the expression for $\nabla a_k(A)$ follows immediately from this. Note that in the last identity we have used that $\mathrm{tr}(AB) = \mathrm{vec}(A^T)^T \cdot \mathrm{vec}(B)$. □

Proposition 9.2 tells us that the variation of the characteristic polynomial of $A \in \mathbb{C}^{n\times n}$ is controlled, to first order, by the trace of $\mathrm{adj}(zI-A)$. This adjugate matrix is an $n\times n$ matrix whose entries are polynomials of degree at most $n-1$ or, equivalently, a matrix polynomial of size $n\times n$ with degree at most $n-1$ (actually, its degree is exactly $n-1$, because of the identity: $(zI-A)\cdot\mathrm{adj}(zI-A) = \det(zI-A)I_n$). Using the explicit expression for the entries of $\mathrm{adj}(zI-A)$, for A being an arbitrary Fiedler matrix M_σ, obtained in Chapter 5, we get one of the main contribution of this chapter, this is, Theorem 9.3.

Theorem 9.3 shows how the coefficients of the characteristic polynomial of any Fiedler companion matrix M_σ change when we perturb M_σ with a dense matrix E. More precisely, we give, to first order in E, the coefficients of the characteristic polynomial of $M_\sigma + E$. Here, the functions $\mathfrak{i}_\sigma(i:j)$ and $\mathfrak{c}_\sigma(i:j)$, and the $n-$tuple EPCIS(σ) (see parts (c) and (d) in Definition 2.8) play an important role.

Theorem 9.3. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\sigma : \{0,1,\ldots,n-1\} \to \{1,\ldots,n\}$ be a bijection with $\mathrm{EPCIS}(\sigma) = (v_0, v_1, \ldots, v_{n-1})$, let M_σ be the Fiedler companion matrix of $p(z)$ associated with σ, and let $E \in \mathbb{C}^{n\times n}$ be an arbitrary matrix. If the characteristic polynomial of $M_\sigma + E$ is denoted by $\widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k$, then, to first order in E,*

$$\widetilde{a}_k - a_k = -\sum_{i,j=1}^{n} p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})\, E_{ij}, \qquad k = 0,1,\ldots,n-1, \tag{9.3}$$

where, for $i,j = 1,2,\ldots,n$, the function $p_{ij}^{(\sigma,k)}(a_0,a_1,\ldots,a_{n-1})$ is a multivariable polynomial in the coefficients of $p(z)$. More precisely, $p_{ij}^{(\sigma,k)}(a_0,a_1,\ldots,a_{n-1})$ is equal to:

(a) *if $v_{n-i} = v_{n-j} = 0$:*

- $a_{k+\mathfrak{i}_\sigma(n-j:n-i)}$,

  if $j \geq i$ *and* $n-k-i+1 \leq \mathfrak{i}_\sigma(n-j:n-i) \leq n-k$;
- $-a_{k+1-\mathfrak{i}_\sigma(n-i:n-j-1)}$,

  if $j < i$ *and* $k+1+i-n \leq \mathfrak{i}_\sigma(n-i:n-j-1) \leq k+1$;
- 0 , *otherwise;*

(b) *if* $v_{n-i} = v_{n-j} = 1$:

- $a_{k+\mathfrak{c}_\sigma(n-i:n-j)}$,

  if $\quad j \le i \quad$ *and* $\quad n-k-j+1 \le \mathfrak{c}_\sigma(n-i:n-j) \le n-k$;
- $-a_{k+1-\mathfrak{c}_\sigma(n-j:n-i-1)}$,

  if $\quad j > i$ *and* $\quad k+1+j-n \le \mathfrak{c}_\sigma(n-j:n-i-1) \le k+1$;
- 0 , *otherwise;*

(c) *if* $v_{n-i} = 1$ *and* $v_{n-j} = 0$:

- 1 , *if* $\quad \mathfrak{i}_\sigma(0:n-j-1) + \mathfrak{c}_\sigma(0:n-i-1) = k$,
- 0 , *otherwise;*

(d) *if* $v_{n-i} = 0$ *and* $v_{n-j} = 1$:

- $\displaystyle\sum_{l=\max\{0,k+1+j-\mathfrak{c}_\sigma(n-j:n-i-1)-n\}}^{l=\min\{k+1-\mathfrak{c}_\sigma(n-j:n-i-1),i-1\}} -\left(a_{n+1-i+l}\, a_{k+1-\mathfrak{c}_\sigma(n-j:n-i-1)-l}\right)$,

  if $\quad j > i$ *and* $\quad k+2+j-i-n \le \mathfrak{c}_\sigma(n-j:n-i-1) \le k+1$;
- $\displaystyle\sum_{l=\max\{0,k+1+i-\mathfrak{i}_\sigma(n-i:n-j-1)-n\}}^{l=\min\{k+1-\mathfrak{i}_\sigma(n-i:n-j-1),j-1\}} -\left(a_{n+1-j+l}\, a_{k+1-\mathfrak{i}_\sigma(n-i:n-j-1)-l}\right)$,

  if $\quad j < i$ *and* $\quad k+2+i-j-n \le \mathfrak{i}_\sigma(n-i:n-j-1) \le k+1$;
- 0 , *otherwise;*

where we set $a_n := 1$.

Proof. From Proposition 9.2, the coefficients of the characteristic polynomial of $M_\sigma + E$ satisfy, to first order in E,

$$\widetilde{a}_k - a_k = -\sum_{i,j=1}^{n} P_{k+1}(j,i) E_{ij},$$

where $P_{k+1}(j,i)$ is the (j,i) entry of P_{k+1} which, according to (9.2) is the kth matrix coefficient of the matrix polynomial $\mathrm{adj}(zI - M_\sigma)$. Therefore $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$ is the kth coefficient of the (j,i) entry of $\mathrm{adj}(zI - M_\sigma)$. From Theorem 5.3 and the proof of Lemma 5.4, we know that the (j,i) entry of $\mathrm{adj}(zI - M_\sigma)$, in each of the cases considered in the statement, is:

(a) $z^{\mathfrak{i}_\sigma(0:n-j-1)+\mathfrak{c}_\sigma(0:n-i-1)} p_{i-1}(z)$, if $j \ge i$, or $z^{\mathfrak{i}_\sigma(n-i+1:n-j-1)} q_{n-i}(z)$, if $j < i$ (see (1) and (2), respectively, in the proof of Lemma 5.4);

(b) $z^{\mathfrak{i}_\sigma(0:n-j-1)+\mathfrak{c}_\sigma(0:n-i-1)} p_{j-1}(z)$, if $j \le i$, or $z^{\mathfrak{c}_\sigma(n-j+1:n-i-1)} q_{n-j}(z)$, if $j > i$ (see (3) and (4) in the proof of Lemma 5.4);

(c) $z^{\mathfrak{i}_\sigma(0:n-j-1)+\mathfrak{c}_\sigma(0:n-i-1)}$ (see (5) in the proof of Lemma 5.4);

(d) $z^{\mathfrak{c}_\sigma(n-j+1:n-i-1)} p_{i-1}(z) q_{n-j}(z)$, if $j > i$, or $z^{\mathfrak{i}_\sigma(n-i+1:n-j-1)} p_{j-1}(z) q_{n-i}(z)$, if $j < i$ (see (6) and (7) in the proof of Lemma 5.4).

Now, it is just a straightforward computation to check that the formulas given in the statement coincide with the kth coefficient of the previous polynomials. □

Remark 9.4. According to the notation in (9.1), we have

$$\nabla a_k(M_\sigma) = -\left[p_{11}^{(\sigma,k)} \ \dots \ p_{n1}^{(\sigma,k)} \ p_{12}^{(\sigma,k)} \ \dots \ p_{n2}^{(\sigma,k)} \ \dots \ p_{1n}^{(\sigma,k)} \ \dots \ p_{nn}^{(\sigma,k)}\right],$$

where we have dropped the dependence of $a_0, \dots, a_{n-1}$ for brevity.

Remark 9.5. For $k = n-1$, and σ an arbitrary bijection, a direct verification in Theorem 9.3 gives

$$p_{ij}^{(\sigma,n-1)}(a_0, \dots, a_{n-1}) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

Then, for any Fiedler matrix M_σ, it follows from (9.3) that

$$a_{n-1}(M_\sigma + E) - a_{n-1}(M_\sigma) = -\sum_{i=1}^{n} E_{ii}.$$

But, since the $(n-1)$th coefficient of the characteristic polynomial of A is equal to $-\mathrm{tr}(A)$, this is a restatement of the well-know identity:

$$\mathrm{tr}(M_\sigma + E) = \mathrm{tr}(M_\sigma) + \mathrm{tr}(E).$$

We want to emphasize that $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ are always linear or quadratic polynomials in the coefficients $a_0, \dots, a_{n-1}$. They depend, at a first stage, on whether the bijection σ has a consecution or an inversion at $n-i$ and $n-j$. In particular, $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ can only be quadratic when there is a consecution at $n-j$ and an inversion at $n-i$. This implies the following corollary.

Corollary 9.6. *Let M_σ be C_1 or C_2 in the statement of Theorem* 9.3*, then $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ in* (9.3) *is a polynomial of degree at most* 1 *in $a_0, \dots, a_{n-1}$, for all $k = 0, 1, \dots, n-1$, and all $1 \le i, j \le n$. For the remaining Fiedler matrices M_σ, there is always some k and some i, j such that $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ is a quadratic polynomial in $a_0, a_1, \dots, a_{n-1}$.*

Proof. Let us first recall that the bijection associated with C_1 is $\sigma_1 = (\sigma_1(0), \sigma_1(1), \dots, \sigma_1(n-1)) = (n, n-1, \dots, 1)$, whereas the bijection associated with C_2 is $\sigma_2 = (\sigma_2(0), \sigma_2(1), \dots, \sigma_2(n-1)) = (1, 2, \dots, n)$ (see Section 2.2). Hence, σ_1 has no consecutions, whereas σ_2 has no inversions.

Then, it remains to show that, if $\sigma : \{0, 1, \dots, n-1\} \to \{1, \dots, n\}$ is a bijection having a consecution at $n-j$ and an inversion at $n-i$, for some $2 \le i, j \le n$, then there is some $0 \le k \le n-1$ such that $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ has degree 2. Note, first, that it must be $i \ne j$. Without loss of generality, let us assume that $j > i$. The proof for the case $j < i$ is analogous. We need to prove that, in the sum defining $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ in the first bullet of case (d) in Theorem 9.3 there is at least one monomial $a_r a_s$ such that $0 \le r, s \le n-1$. More precisely, we need to prove:

(i) There is some $0 \le k \le n-1$ such that $k + 2 + j - i - n \le \mathfrak{c}_\sigma(n-j : n-i-1) \le k+1$.

(ii) There is some l, with $\max\{0, k+1+j-\mathfrak{c}_\sigma(n-j : n-i-1)-n\} \le l \le \min\{k+1-\mathfrak{c}_\sigma(n-j : n-i-1), i-1\}$, such that $0 \le n+1-i+l \le n-1$ and $0 \le k+1-\mathfrak{c}_\sigma(n-j : n-i-1)-l \le n-1$.

For this, it suffices to take $k = \mathfrak{c}_\sigma(n-j : n-i-1) - 1 = \mathfrak{c}_\sigma(n-j+1 : n-i-1)$ and $l = 0$. Note that (ii) is fulfilled for these values of k and l, because $i \ge 2$. □

The expressions given in Theorem 9.3 for the variation of the coefficients of the characteristic polynomial of M_σ are involved in general (that is, for arbitrary Fiedler matrices). We will show them explicitly in Section 9.1.2 for some Fiedler matrices, including the Frobenius companion matrices.

The following result describes some properties of the polynomials $p_{ij}^{(\sigma,k)}(a_0, \dots, a_{n-1})$ that will be used later.

Lemma 9.7. *Let $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$ be the polynomial defined in* (9.3)*. Then:*

(a) *For $k = 0, 1, \ldots, n-1$,*

$$p_{ii}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = \begin{cases} a_{k+1} & \text{if} \quad i \geq n-k\,, \\ 0 & \text{if} \quad i < n-k\,, \end{cases}$$

with $a_n = 1$.

(b) *If σ has a consecution at $n-2$, then $p_{12}^{(\sigma,0)}(a_0, a_1, \ldots, a_{n-1}) = -a_0$, and if σ has an inversion at $n-2$, then $p_{21}^{(\sigma,0)}(a_0, a_1, \ldots, a_{n-1}) = -a_0$.*

Proof. From Theorem 9.3 we have $p_{ii}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = a_{k+1}$, if $n-1 \geq k \geq n-i$ (namely, if $i \geq n-k$), and $p_{ii}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = 0$ otherwise. This proves part (a).

For part (b), if σ has a consecution at $n-2$, then following the notation of Theorem 9.3, we have $v_{n-2} = v_{n-1} = 1$, and $\mathfrak{c}_\sigma(n-2 : n-2) = 1$, so part (b) of Theorem 9.3 gives $p_{12}^{(\sigma,0)}(a_0, a_1, \ldots, a_{n-1}) = -a_0$. Similarly, if σ has an inversion at $n-2$, then $v_{n-2} = v_{n-1} = 0$, and part (a) of Theorem 9.3 gives $p_{21}^{(\sigma,0)}(a_0, a_1, \ldots, a_{n-1}) = -a_0$. □

To identify those indices k for which $\nabla a_k(M_\sigma)$ contains quadratic terms in $a_0, \ldots, a_{n-1}$ may be interesting in practice. Notice that the presence of such quadratic terms implies that the sensitivity of the coefficient $a_k(M_\sigma)$ to perturbations of M_σ is quadratic in $a_0, \ldots, a_{n-1}$, instead of linear. This implies in turn that, for large values of $a_0, \ldots, a_{n-1}$, we can expect much larger changes after small perturbations in these coefficients than in the ones where $\nabla a_k(M_\sigma)$ contains only linear terms. We have seen in Corollary 9.6 that, for all Fiedler matrices but the Frobenius ones, there is always at least one k such that $\nabla a_k(M_\sigma)$ contains quadratic terms. Moreover, the proof of Corollary 9.6 tells us that if i, j are such that σ has a consecution at $n-j$ and an inversion at $n-i$, and $j > i$ (respectively, $j < i$), then for $k = \mathfrak{c}_\sigma(n-j+1 : n-i-1)$ (resp., $k = \mathfrak{i}_\sigma(n-i+1 : n-j-1)$) the gradient $\nabla a_k(M_\sigma)$ contains quadratic terms. In particular, Lemma 9.8 states that, for all Fiedler matrices but the Frobenius ones, $\nabla a_0(M_\sigma)$ contains always quadratic polynomials in $a_0, \ldots, a_{n-1}$

Lemma 9.8. *Let $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$ be the polynomial defined in* (9.3)*, and let $t \in \{0, 1, \ldots, n-3\}$.*

(a) *If $PCIS(\sigma) = (v_0, v_1, \ldots, v_t = 1, v_{t+1} = 0, v_{t+2} = 0, \ldots, v_{n-2} = 0)$ then*

$$p_{2,n-t}^{(\sigma,0)}(a_0, a_1, \ldots, a_{n-1}) = -a_{n-1}a_0.$$

(b) *If $PCIS(\sigma) = (v_0, v_1, \ldots, v_t = 0, v_{t+1} = 1, v_{t+2} = 1, \ldots, v_{n-2} = 1)$ then*

$$p_{n-t,2}^{(\sigma,0)}(a_0, a_1, \ldots, a_{n-1}) = -a_{n-1}a_0.$$

Proof. We prove only part (a) because part (b) is similar. Since $n - t > 2$ and $k - \mathfrak{c}_\sigma(n-j+1 : n-i-1) = -\mathfrak{c}_\sigma(t+1 : n-3) = 0$, from part (d) of Theorem 9.3,we have

$$p_{2,n-t}^{(\sigma,0)}(a_0, a_1, \ldots, a_{n-1}) = \sum_{l=\max\{0,-\mathfrak{c}_\sigma(t+1:n-3)-t\}}^{l=\min\{0,1\}} -a_{n-1+l}\; a_{k-\mathfrak{c}_\sigma(t+1:n-3)-l} = -a_{n-1}a_0.$$

□

The main result, from the practical point of view, in this section is a direct consequence of Theorem 9.3.

Corollary 9.9. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial, and M_σ be a Fiedler companion matrix of $p(z)$. Assume that the roots of $p(z)$ are computed as the eigenvalues of M_σ with a backward stable algorithm, i.e., an algorithm that computes the exact eigenvalues of some matrix $M_\sigma + E$, with $\|E\|_\infty = O(u)\|M_\sigma\|_\infty$. Then the computed roots are the exact roots of a polynomial $\widetilde{p}(z)$ such that:*

(a) *If $M_\sigma = C_1, C_2$,*

$$\frac{\|\widetilde{p} - p\|_\infty}{\|p\|_\infty} = O(u)\|p\|_\infty, \tag{9.4}$$

(b) *if $M_\sigma \neq C_1, C_2$,*

$$\frac{\|\widetilde{p} - p\|_\infty}{\|p\|_\infty} = O(u)\|p\|_\infty^2, \tag{9.5}$$

where u is the machine precision. In other words, the backward error of the computed roots $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$ is

$$\eta_\infty(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n) = \begin{cases} O(u)\|p\|_\infty, & \text{if } M_\sigma = C_1, C_2, \\ O(u)\|p\|_\infty^2, & \text{if } M_\sigma \neq C_1, C_2. \end{cases}$$

Proof. If the eigenvalues of M_σ are computed with a backward stable algorithm, the computed eigenvalues are the exact eigenvalues of a matrix $M_\sigma + E$, for some matrix $E \in \mathbb{C}^{n\times n}$ such that $\|E\|_\infty = O(u)\|M_\sigma\|_\infty$. Thus, the computed eigenvalues are the exact roots of the polynomial $\widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k = \det(zI - M_\sigma - E)$. From Theorem 9.3, to first order in E,

$$\begin{aligned} |\widetilde{a}_k - a_k| &= \left| \sum_{i,j=1}^{n} p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) E_{ij} \right| \leq \sum_{i,j=1}^{n} \left| p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) \right| \cdot |E_{ij}| \\ &\leq \left(\max_{1\leq i,j\leq n} |E_{ij}| \right) \cdot \left(\sum_{i,j=1}^{n} |p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})| \right). \end{aligned}$$

Notice, also from Theorem 9.3, that the absolute value of every polynomial $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$ is bounded by $n\|p\|_\infty^2$ and that, by Corollary 9.6, the square in the norm of p is necessary in all Fiedler matrices except the Frobenius companion matrices, where it can be replaced by 1. Therefore,

$$\max_{k=0,1,\ldots,n-1} |\widetilde{a}_k - a_k| = \|\widetilde{p} - p\|_\infty = O(u)\|M_\sigma\|_\infty \|p\|_\infty^2 = O(u)\|p\|_\infty^3,$$

where we have used that $\max_{i,j=1,2,\ldots,n} |E_{ij}| = O(u)\|M_\sigma\|_\infty$ and $\|M_\sigma\|_\infty = O(1)\|p\|_\infty$ (see [50, Th. 3.3]). □

Remark 9.10. Note that Corollary 9.9 implies that computing the roots of $p(z)$ using any of the Fiedler matrices of $p(z)$ is not backward stable if $\|p\|_\infty$ is large. For the Frobenius companion matrices, Corollary 9.9 recovers (1.13).

As a consequence of (9.4) and (9.5) we get the following conclusions:

(C1) From the point of view of the normwise backward errors in the (monic) polynomial $p(z)$, any Fiedler matrix can be used for solving the root-finding problem with the same reliability as Frobenius companion matrices when $\|p\|_\infty = O(1)$. In this case, the root-finding problem solved by applying a backward stable eigenvalue algorithm on any Fiedler companion matrix is a backward stable method from the polynomial point of view.

(C2) However, when $\|p\|_\infty$ is large none of the Fiedler matrices leads to a backward stable algorithm for the root-finding problem and, moreover, any Fiedler matrix other than Frobenius companion matrices may produce much larger backward errors than the ones produced when using Frobenius matrices.

Note, in particular, that since $\|p\|_\infty \geq 1$, no Fiedler matrix can improve the behavior of the bounds of Frobenius matrices in the root-finding problem from the point of view of backward errors.

It is worth to remark that if the matrix E in the statement of Corollary 9.9 satisfies $\|E\|_\infty = c(p)O(u)\|M_\sigma\|_\infty$, with $c(p)$ being some positive quantity depending on $p(z)$ then, with the appropriate changes in the proof of Corollary 9.9, we could replace (9.4) and (9.5) by, respectively:

$$\frac{\|\widetilde{p}-p\|_\infty}{\|p\|_\infty} = c(p)O(u)\|p\|_\infty \qquad \text{and} \qquad \frac{\|\widetilde{p}-p\|_\infty}{\|p\|_\infty} = c(p)O(u)\|p\|_\infty^2.$$

Hence, even for eigensolvers whose backward stability can not be guaranteed (like the fast QR-like algorithms mentioned in the Introduction for the Frobenius companion matrix or those that can be applied to other Fiedler matrices) our developments allow us to provide backward error estimates for the polynomial root-finding problem using Fiedler companion matrices.

9.1.1 Recursive formula for the derivatives of the characteristic polynomial

In Chapter 5 we have given an explicit formula for the entries of $\operatorname{adj}(zI - M_\sigma)$, with M_σ being an arbitrary Fiedler matrix. The aim of this subsection is to provide, in Proposition 9.11, a recursive formula for the coefficients of $\operatorname{adj}(zI - A)$ when viewed as a matrix polynomial in z, where $A \in \mathbb{C}^{n\times n}$ is an arbitrary matrix. This is an interesting theoretical result that gives an alternative description of the coefficients of $\operatorname{adj}(zI - A)$ and, as a consequence of Lemma 9.1, of the gradient of the characteristic polynomial of A. But it may also have a practical interest, as it provides a recursive way to construct these coefficients.

Proposition 9.11. [66, Ch. 4, §4] *Let $A \in \mathbb{C}^{n\times n}$, and let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be the characteristic polynomial of A. Let the matrices $A_1, A_2, \ldots, A_n \in \mathbb{C}^{n\times n}$ be defined by the following recurrence relation*

$$\begin{cases} A_n = I\,, \quad \textit{and} \\ A_k = A \cdot A_{k+1} + a_k I, \quad \textit{for } k = n-1, n-2, \ldots, 1. \end{cases} \tag{9.6}$$

Then,

$$\operatorname{adj}(zI - A) = \sum_{k=0}^{n-1} z^k A_{k+1}.$$

We note that, as a consequence of the recursive relations of the Horner shifts (2.6), the matrices A_k are the Horner shifts of $p(z) = \det(zI - A)$ evaluated at A. More precisely:

$$A_k = p_{n-k}(A) = A^{n-k} + a_{n-1}A^{n-k-1} + \cdots + a_{k+1}A + a_k I\,.$$

With this in mind, Proposition 9.2 gives the following expression for the gradient of the kth coefficient of the characteristic polynomial of A:

$$\nabla a_k(A) = -\left[\operatorname{vec}(p_{n-k-1}(A^T))\right]^T, \quad \text{for} \quad k = 0, 1, \ldots, n-1. \tag{9.7}$$

Proposition 9.11 has been used in [53] to get an explicit formula for the derivatives of the coefficients of $\det(zI - C)$, with C being a Frobenius companion matrix. For this, the authors

take advantage of the explicit expression of the matrices A_k defined in (9.6) with $A = C$, which are very simple in this case (see [53, p. 768]). However, for A being an arbitrary Fiedler matrix, the matrices A_k become much more involved, and it is not easy to get an explicit expression of these matrices just by using (9.6). For this reason, we have obtained the expression of the entries of $\operatorname{adj}(zI - A)$ by other means. However, Proposition 9.11 gives us an alternative way to get $\operatorname{adj}(zI - A)$ using the Horner shifts of A.

We want to emphasize that, as a consequence of the previous remarks, the polynomial $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$ in Theorem 9.3 corresponds to the (j, i) entry of the matrix $p_{n-k-1}(M_\sigma)$. In the following section, we display these matrices for some particular relevant cases, including the Frobenius companion matrices. It is also interesting to note that Corollary 9.6 implies that the first and second Frobenius companion matrices are the only Fiedler matrices M_σ for which all Horner shifts $p_k(M_\sigma)$ have entries which are linear multivariable polynomials in the coefficients of $p(z)$. For all other Fiedler matrices M_σ, there is at least one k such that $p_k(M_\sigma)$ contains some quadratic entries.

9.1.2 Some particular cases

We obtain in this section the explicit expression (9.3) for particular Fiedler matrices that are, or may be, of interest in practice. We start with the classical Frobenius companion matrices in Theorem 9.12, where we get analogous formulas to the ones obtained in [53] for the Frobenius companion matrix considered in that paper.

Theorem 9.12. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $C = C_1$ or C_2 be the first or second Frobenius companion matrix of $p(z)$, and let $E \in \mathbb{C}^{n\times n}$. If $\widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k$ is the characteristic polynomial of $C + E$. Then, to first order in E, for $k = 0, 1, \ldots, n-1$:*

(i) *If $C = C_1$:*

$$\widetilde{a}_k - a_k = \sum_{s=0}^{k} \sum_{j=1}^{n-k-1} a_s E_{j-s+k+1,j} - \sum_{s=k+1}^{n} \sum_{j=n-k}^{n} a_s E_{j-s+k+1,j}. \tag{9.8}$$

(ii) *If $C = C_2$:*

$$\widetilde{a}_k - a_k = \sum_{s=0}^{k} \sum_{i=1}^{n-k-1} a_s E_{i,i-s+k+1} - \sum_{s=k+1}^{n} \sum_{i=n-k}^{n} a_s E_{i,i-s+k+1}. \tag{9.9}$$

Proof. For claim (i), we recall that, if σ is the bijection associated with C_1, then $\operatorname{PCIS}(\sigma) = (0, \ldots, 0)$. For this bijection, $\mathfrak{i}_\sigma(n - j : n - i) = j - i + 1$ holds for $i \le j$. Then, applying part (a) in Theorem 9.3, we get

$$\widetilde{a}_k - a_k = \sum_{\substack{j<i \\ k+1+i-n\le i-j\le k+1}} a_{j-i+1+k} E_{ij} - \sum_{\substack{j\ge i \\ n-k-i+1\le j-i+1\le n-k}} a_{j-i+1+k} E_{ij}.$$

With the change of variables $s = j - i + 1 + k$ the claim is proved.

For claim (ii), we recall that the bijection σ associated with C_2 satisfies $\operatorname{PCIS}(\sigma) = (1, \ldots, 1)$. For this bijection, $\mathfrak{c}_\sigma(n - i : n - j) = i - j + 1$, when $j \le i$. Then, applying part (b) in Theorem 9.3, we get

$$\widetilde{a}_k - a_k = \sum_{\substack{j>i \\ k+1+j-n\le j-i\le k+1}} a_{i-j+1+k} E_{ij} - \sum_{\substack{j\le i \\ n-k-j+1\le i-j+1\le n-k}} a_{i-j+1+k} E_{ij}.$$

Again, we use the change of variables $s = i - j + 1 + k$ to get the result . □

According to (9.7), the matrix $p_{n-k-1}(A^T)$ encodes the information about $\nabla a_k(A)$. In the case of Frobenius companion matrices, these Horner shifts can be computed without too much effort, since they are equal to:

$$p_{n-k-1}(C_1^T) = p_{n-k-1}(C_2) = \left[\begin{array}{ccc|cccc} 0 & \dots & 0 & 1 & & & 0 \\ -a_k & & & a_{n-1} & 1 & & \\ \vdots & \ddots & & \vdots & a_{n-1} & \ddots & \\ -a_1 & \ddots & -a_k & a_{k+1} & \vdots & \ddots & 1 \\ -a_0 & \ddots & \vdots & & a_{k+1} & \ddots & a_{n-1} \\ & \ddots & -a_1 & & & \ddots & \vdots \\ 0 & & -a_0 & 0 & & & a_{k+1} \end{array}\right], \tag{9.10}$$

for $k = 0, 1, \dots, n-1$, where the first block-column contains $n-k-1$ columns, and the second block-column contains $k+1$ columns. The reader may check that, indeed, the (i, j) entry of (9.10) is the coefficient of E_{ij} in (9.8). The same happens with the transpose of (9.10) and formula (9.9).

Excluding the Frobenius companion matrices, the simplest Fiedler matrices are F (defined in (2.8)), and F^T, with just one inversion (resp., consecution) at 0, and consecutions (resp., inversions) elsewhere.

Theorem 9.13. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $F = M_\sigma$ be the Fiedler companion matrix of $p(z)$ with PCIS$(\sigma) = (0, 1, 1, , \dots, 1)$ and let $E \in \mathbb{C}^{n\times n}$. If $\widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k$ is the characteristic polynomial of $F + E$, then, to first order in E,*

$$\begin{aligned} \widetilde{a}_k - a_k = & \sum_{j=k+1}^{n-1} a_0 a_{n+k+1-j} E_{nj} + \sum_{s=0}^{k} \sum_{i=1}^{n-k-2} a_s E_{i,i+k+1-s} + \sum_{s=1}^{k} a_s E_{n-k-1,n-s} - E_{n-k-1,n} \\ & - \sum_{s=k+1}^{n} \sum_{i=n-k}^{n-1} a_s E_{i,i+k+1-s} - E_{n-k-1,n} - a_{k+1} E_{nn}. \end{aligned} \tag{9.11}$$

Proof. To compute the polynomials $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ in (9.3), according to Theorem 9.3, we distinguish the following cases:

(a) If $j = n$ and $i = 1, 2, \dots, n-1$, we have $v_{n-j} = 0$ and $v_{n-i} = 1$ and $\mathfrak{i}_\sigma(0 : n-j-1) + \mathfrak{c}_\sigma(0 : n-i-1) = n-i-1$. Therefore, $p_{in}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = 1$ if $i = n-k-1$ and $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = 0$ otherwise.

(b) If $j = n$ and $i = n$, we have $v_{n-j} = v_{n-i} = 0$ and $\mathfrak{i}_\sigma(n-j : n-i-1) = 0$. Therefore, $p_{nn}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = a_{k+1}$ if $n-1 \geq k \geq 0$ and $p_{nn}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = 0$ otherwise.

(c) If $j = 1, 2, \dots, n-1$ and $i = n$, we have $v_{n-j} = 1$ and $v_{n-i} = 0$ and $\mathfrak{i}_\sigma(n-i+1 : n-j-1) = 0$. Therefore, $p_{nj}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = -a_{n+k-j+1} a_0$ if $j \geq k+1$ and $p_{nj}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = 0$ if $j < k+1$.

(d) If $i, j = 1, 2, \dots, n-1$ and $j \leq i$, we have $v_{n-i} = v_{n-j} = 1$ and $\mathfrak{c}_\sigma(n-i : n-j) = i-j+1$. Therefore, $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = a_{k+1+i-j}$ if $n-k-j+1 \leq i-j+1 \leq n-k$ and $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = 0$ otherwise. With the change of variable $s = k+1+i-j$ we get $p_{i,i+k+1-s}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = a_s$ if $k+1 \leq s \leq n$ and $n-k \leq i \leq n-1$, and $p_{i,i+k+1-s}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1}) = 0$ otherwise.

(e) If $i, j = 1, 2, \ldots, n-1$ and $j > i$, we have $v_{n-i} = v_{n-j} = 1$ and $\mathfrak{c}_\sigma(n-j : n-i-1) = j - i$. Therefore, $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = a_{k+1+i-j}$ if $k + 1 + j - n \leq j - i \leq k + 1$ and $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = 0$ otherwise. With the change of variable $s = k + 1 + i - j$ we get $p_{i,i+k+1-s}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = a_s$ if $0 \leq s \leq k$ and $1 \leq i \leq n - k - 1$, and $p_{i,i+k+1-s}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = 0$ otherwise.

□

Theorem 9.13 illustrates how a single change in the PCIS of the Frobenius companion matrix, i.e., just to change the position of the factor M_0 in the product defining C_1 and C_2, implies the appearance of quadratic terms in the formula for the gradient of the coefficients of the characteristic polynomial (see the first summand in the right-hand-side of (9.11)). As before, this can also be seen by explicitly displaying the Horner shifts evaluated at F:

$$p_{n-k-1}(F) = \left[\begin{array}{ccccc|ccccc} 0 & & & & & 1 & & & & 0 \\ -a_k & & & & & a_{n-1} & \ddots & & & \vdots \\ \vdots & \ddots & & & & \vdots & \ddots & 1 & & 0 \\ -a_1 & & -a_k & & & a_{k+2} & & a_{n-1} & & -a_0 \\ -a_0 & \ddots & \vdots & -a_k & & a_{k+1} & \ddots & \vdots & & -a_0a_{n-1} \\ & \ddots & -a_1 & \vdots & & & \ddots & a_{k+2} & & \vdots \\ & & -a_0 & -a_1 & & & & a_{k+1} & & -a_0a_{k+2} \\ & & & 1 & & & & & & a_{k+1} \end{array}\right],$$

for $k = 0, 1, \ldots, n-3$,

$$p_1(F) = \begin{bmatrix} 0 & & & & & & 0 \\ -a_{n-2} & 1 & & & & & \\ -a_{n-3} & a_{n-1} & 1 & & & & \\ \vdots & & a_{n-1} & \ddots & & & \\ \vdots & & & \ddots & 1 & & \\ -a_1 & & & & a_{n-1} & -a_0 & \\ 1 & & & & 0 & a_{n-1} & \end{bmatrix}, \qquad \text{and } p_0(F) = I.$$

The number of columns in the first block-column of $p_{n-k-1}(F)$ above is $n-k-1$, and the number of columns in the second block column is $k+1$. The reader may check that the (i, j) entry of $p_{n-k-1}(F)^T$ is the coefficient of E_{ij} in (9.11).

Our last example is the case of the pentadiagonal Fiedler matrix P_1 in (2.7). Formulas here, as can be seen in Theorem 9.14, become much more involved.

Theorem 9.14. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $P_1 = M_\sigma$ be the Fiedler companion matrix of $p(z)$ with PCIS$(\sigma) = (1, 0, 1, 0, \ldots)$ and let $E \in \mathbb{C}^{n\times n}$. If $\widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k$ is the characteristic polynomial of $P_1 + E$, then, to first order in E,*

$$\widetilde{a}_k - a_k = -\sum_{s=k+1}^{n} \left(\sum_{r=\lceil \frac{n+s}{2} \rceil - k}^{n/2} a_s E_{2(k+r-s)+1,2r-1} + \sum_{r=\lfloor \frac{n+s}{2} \rfloor - k}^{n/2} a_s E_{2r,2(k+r-s+1)} \right)$$
$$+ \sum_{s=0}^{k} \left(\sum_{r=1}^{\lceil \frac{n+s}{2} \rceil - k - 1} a_s E_{2(k+r-s)+1,2r-1} + \sum_{r=1}^{\lfloor \frac{n+s}{2} \rfloor - k - 1} a_s E_{2r,2(k+r-s+1)} \right) - \sum_{s=\max\{1, \frac{n}{2}-k\}}^{\min\{\frac{n}{2}, n-k-1\}} E_{2s,2(n-k-s)-1}$$

$$+\sum_{s=\max\{0,2k-n+2\}}^{k}\left(\sum_{r=1}^{s-k+\frac{n}{2}}\sum_{m=\max\{0,2k+2r-s-n\}}^{\min\{s,2r-2\}}a_{n-2r+2+m}a_{s-m}E_{2r-1,2(k-s+r)}\right.$$
$$\left.+\sum_{r=1}^{s-k+\frac{n}{2}-1}\sum_{m=\max\{0,2k+2r-s-n+1\}}^{\min\{s,2r-1\}}a_{n-2r+1+m}a_{s-m}E_{2(k+r-s)+1,2r}\right)$$

if n is an even number, or

$$\widetilde{a}_k-a_k=-\sum_{s=k+1}^{n}\left(\sum_{r=\lceil\frac{n+s}{2}\rceil-k}^{\frac{n+1}{2}}a_sE_{2r-1,2(k-s+r)+1}+\sum_{r=\lfloor\frac{n+s}{2}\rfloor-k}^{\frac{n-1}{2}}a_sE_{2(k-s+r+1),2r}\right)$$
$$+\sum_{s=0}^{k}\left(\sum_{r=1}^{\lceil\frac{n+s}{2}\rceil-k-1}a_sE_{2r-1,2(k-s+r)+1}+\sum_{r=1}^{\lfloor\frac{n+s}{2}\rfloor-k-1}a_sE_{2(k-s+r+1),2r}\right)-\sum_{s=\max\{1,\frac{n+1}{2}-k\}}^{\min\{\frac{n+1}{2},n-k-1\}}E_{2s-1,2(n-k-s)}$$
$$+\sum_{s=\max\{0,2k-n+2\}}^{k}\sum_{r=1}^{s-k+\frac{n-1}{2}}\left(\sum_{m=\max\{0,2k-n-s+2r+1\}}^{\min\{s,2r-1\}}a_{n-2r+1+m}a_{s-m}E_{2r,2(k-s+r)+1}\right.$$
$$\left.+\sum_{m=\max\{0,2k-n-s+2r\}}^{\min\{s,2r-2\}}a_{n-2r+2+m}a_{s-m}E_{2(k-s+r),2r-1}\right)$$

if n is an odd number.

Proof. We give a sketch of the proof when the degree of $p(z)$ is even, since the odd degree case is similar. To compute the polynomials $p_{ij}^{(\sigma,k)}(a_0,a_1,\ldots,a_{n-1})$ in (9.3) we have to distinguish several cases.

(a) If i and j are odd numbers, then we get

$$p_{ij}^{(\sigma,k)}(a_0,a_1,\ldots,a_{n-1})=\begin{cases}a_{k+1+\frac{j-i}{2}}, & \text{if } j\geq i \text{ and } n-\frac{i+j}{2}\leq k\leq n-1-\frac{j-i}{2},\\ -a_{k+1+\frac{j-i}{2}}, & \text{if } j<i \text{ and } \frac{i-j}{2}-1\leq k\leq n-1-\frac{i+j}{2},\\ 0, & \text{otherwise.}\end{cases}$$

(b) If i and j are even numbers, then we get

$$p_{ij}^{(\sigma,k)}(a_0,a_1,\ldots,a_{n-1})=\begin{cases}a_{k+1+\frac{i-j}{2}}, & \text{if } j\leq i \text{ and } n-\frac{i+j}{2}\leq k\leq n-1-\frac{i-j}{2},\\ -a_{k+1+\frac{i-j}{2}}, & \text{if } j>i \text{ and } \frac{j-i}{2}-1\leq k\leq n-1-\frac{i+j}{2} \quad \text{and}\\ 0, & \text{otherwise.}\end{cases}$$

(c) If i is an odd number and j is an even number, then we get

$$p_{ij}^{(\sigma,k)}(a_0,a_1,\ldots,a_{n-1})=\begin{cases}1, & \text{if } k=n-\frac{i+j+1}{2},\\ 0, & \text{otherwise.}\end{cases}$$

(d) If i is an even number and j is an odd number, then we get

$$p_{ij}^{(\sigma,k)}(a_0,a_1,\ldots,a_{n-1})$$
$$=\begin{cases}-\sum_{m=\max\{0,k-\frac{j-i-1}{2}-n+j\}}^{\min\{k-\frac{j-i-1}{2},i-1\}}a_{n-i+1+m}a_{k-\frac{j-i-1}{2}-m}, & \text{if } j>i \text{ and } \frac{j-i-1}{2}\leq k\leq n-\frac{j-i+3}{2},\\ -\sum_{m=\max\{0,k-\frac{i-j-1}{2}-n+i\}}^{\min\{k-\frac{i-j-1}{2},j-1\}}a_{n-j+1+m}a_{k-\frac{i-j-1}{2}-m}, & \text{if } j<i \text{ and } \frac{i-j-1}{2}\leq k\leq n-\frac{i-j+3}{2},\\ 0, & \text{otherwise.}\end{cases}$$

The result follows from these formulas for $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$, together with some algebraic manipulations and appropriate changes of variables. □

For the pentadiagonal Fiedler matrix P_1, the matrices $p_{n-k-1}(P_1)$, for $k = 0, 1, \ldots, n-1$, do not have a simple structure. For illustrative purposes, we include here a 6×6 example. Let P_1 be the pentadiagonal Fiedler matrix, in (2.7), of the polynomial $p(z) = z^6 + \sum_{k=0}^{5} a_k z^5$. Then, it can be seen that

$$p_0(P_1) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad p_1(P_1) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -a_4 & a_5 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & a_5 & 0 & 0 & 0 \\ 0 & 0 & -a_2 & a_5 & -a_1 & 1 \\ 0 & 0 & 1 & 0 & a_5 & 0 \\ 0 & 0 & 0 & 0 & -a_0 & a_5 \end{bmatrix},$$

$$p_2(P_1) = \begin{bmatrix} 0 & 0 & -a_3 & 1 & 0 & 0 \\ -a_3 & 0 & -a_2 - a_3 a_5 & a_5 & -a_1 & 1 \\ 0 & 1 & a_4 & 0 & 0 & 0 \\ -a_2 & 0 & -a_1 - a_2 a_5 & a_4 & -a_0 - a_1 a_5 & a_5 \\ 1 & 0 & a_5 & 0 & a_4 & 0 \\ 0 & 0 & -a_0 & 0 & -a_0 a_5 & a_4 \end{bmatrix},$$

$$p_3(P_1) = \begin{bmatrix} 0 & 0 & -a_2 & 0 & -a_1 & 1 \\ -a_2 & 0 & -a_1 - a_2 a_5 & 0 & -a_0 - a_1 a_5 & a_5 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -a_1 & -a_2 & -a_0 - a_1 a_5 - a_2 a_4 & a_3 & -a_0 a_5 - a_1 a_4 & a_4 \\ 0 & 1 & a_4 & 0 & a_3 & 0 \\ -a_0 & 0 & -a_0 a_5 & 0 & -a_0 a_4 & a_3 \end{bmatrix},$$

$$p_4(P_1) = \begin{bmatrix} 0 & 0 & -a_1 & 0 & -a_0 & 0 \\ -a_1 & 0 & -a_0 - a_1 a_5 & 0 & -a_0 a_5 & 0 \\ 0 & 0 & 0 & 0 & -a_1 & 1 \\ -a_0 & -a_1 & -a_0 a_5 - a_1 a_4 & 0 & -a_0 a_4 - a_1 a_3 & a_3 \\ 0 & 0 & 0 & 1 & a_2 & 0 \\ 0 & -a_0 & -a_0 a_4 & 0 & -a_0 a_3 & a_2 \end{bmatrix},$$

$$p_5(P_1) = \begin{bmatrix} 0 & 0 & -a_0 & 0 & 0 & 0 \\ -a_0 & 0 & -a_0 a_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -a_0 & 0 \\ 0 & -a_0 & -a_0 a_4 & 0 & -a_0 a_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -a_0 & -a_0 a_2 & a_1 \end{bmatrix}.$$

Unlike the previous cases C_1, C_2 and F, there does not seem to be a simple pattern for $p_{n-k-1}(P_1)$ for arbitrary n.

9.1.3 Balancing and backward error

Balancing first computes in exact arithmetic a matrix $DM_\sigma D^{-1}$, where D is diagonal, which has the same characteristic polynomial as M_σ (see Section 1.2.3). Then a backward stable algorithm is applied to compute the eigenvalues of $DM_\sigma D^{-1}$, so that we get the exact eigenvalues of $DM_\sigma D^{-1} + \widetilde{E}$, with

$$\|\widetilde{E}\| = O(u)\|DM_\sigma D^{-1}\|, \tag{9.12}$$

for some matrix norm $\|\cdot\|$. Now, we can get a formula like (9.3) for the change of the coefficients of the characteristic polynomial of $DM_\sigma D^{-1}$ using the identity:

$$\det(zI - DM_\sigma D^{-1} - \widetilde{E}) = \det(zI - M_\sigma - D^{-1}\widetilde{E}D),$$

and applying Theorem 9.3 with the perturbation $D^{-1}\widetilde{E}D$ instead of E. In particular, following the arguments in the proof of Corollary 9.9, we get

$$|\widetilde{a}_k - a_k| \leq n^2 \max_{1\leq i,j\leq n} \left(\left| p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) \, \frac{d_j}{d_i} \right| \right) \cdot \max_{1\leq i,j\leq n} |\widetilde{E}_{ij}| \,,$$

with $\widetilde{E}$ as in (9.12). In this way, we get a formula which provides an "a posteriori" (that is, once the diagonal parameters d_i are known) measure for the backward error of the polynomial root-finding problem using balanced Fiedler matrices.

Though the numerical experiments carried out in Section 9.4 indicate that balancing usually produces smaller backward errors, we will see in Proposition 9.20 that, for any degree, there are infinitely many polynomials for which the condition numbers of all coefficients of the characteristic polynomial of any matrix $DM_\sigma D^{-1}$ are large. This shows that, though in practice balancing Fiedler matrices may be a good strategy for the root-finding problem, there are polynomials with arbitrary degree, for which the strategy does not lead to small backward errors.

9.2 Conditioning of the characteristic polynomial

The developments carried out in Section 9.1 are closely related to the conditioning of the characteristic polynomial of the matrix M_σ. The condition number of the characteristic polynomial provides a measure of its sensitivity to perturbations of the matrix. As we have seen, this is in turn related with the gradient of the coefficients of the characteristic polynomial. In this section, we introduce the condition number (absolute and relative) for the coefficients of the characteristic polynomial, and we relate it with (the norm of) its gradient. In this way, we will see that, from the polynomial point of view, the backward stability of the polynomial root-finding problem via eigenvalue methods is determined by the conditioning of the characteristic polynomial.

For a given matrix $A \in \mathbb{C}^{n\times n}$, let us first assume that the entries of the matrix E in (1.8) satisfy $|E_{ij}| \leq \epsilon \|\mathrm{vec}(A)\|_\infty$. Then, using Holder's inequality $|u^T v| \leq \|u^T\|_\infty \|v\|_\infty$ (with $\| \begin{bmatrix} u_1 & \ldots & u_n \end{bmatrix} \|_\infty = |u_1| + \cdots + |u_n|$)[1], from (1.8) we get, up to first order, the following inequalities:

$$\begin{aligned} |a_k(A+E) - a_k(A)| = |\nabla a_k(A) \cdot \mathrm{vec}(E)| &\leq \|\nabla a_k(A)\|_\infty \|\mathrm{vec}(E)\|_\infty \\ &\leq \epsilon \|\nabla a_k(A)\|_\infty \|\mathrm{vec}(A)\|_\infty \,. \end{aligned} \tag{9.13}$$

It is straightforward to show that there exists a particular matrix E with $\|\mathrm{vec}(E)\|_\infty = \epsilon \|\mathrm{vec}(A)\|_\infty$ such that $|\nabla a_k(A) \cdot \mathrm{vec}(E)| = \|\nabla a_k(A)\|_\infty \|\mathrm{vec}(E)\|_\infty$. For this matrix the bound in (9.13) is attained to first order in ϵ. With this in mind, Proposition 9.15 immediately follows.

Proposition 9.15. *Let $A \in \mathbb{C}^{n\times n}$ and $a_k(X) : \mathbb{C}^{n^2} \to \mathbb{C}$ be the kth coefficient of the characteristic polynomial of $X \in \mathbb{C}^{n\times n}$, considered as a function of X. We define the condition numbers $\kappa(a_k, A)$ and $\kappa_{\mathrm{rel}}(a_k, A)$ as*

$$\kappa(a_k, A) := \lim_{\epsilon\to 0} \sup \left\{ \frac{|a_k(A+E) - a_k(A)|}{\epsilon} : \|\mathrm{vec}(E)\|_\infty \leq \epsilon \|\mathrm{vec}(A)\|_\infty \right\} \tag{9.14}$$

[1]Note that, according to the definition of $\|\cdot\|_\infty$ for $m \times n$ matrices, see [79, p. 108], the expressions for $\|u\|_\infty$ and $\|u^T\|_\infty$, i.e., for column and row vectors, are different.

and

$$\kappa_{\rm rel}(a_k, A) := \lim_{\epsilon\to 0}\sup\left\{\frac{|a_k(A+E)-a_k(A)|}{\epsilon|a_k(A)|} : \|\mathrm{vec}(E)\|_\infty \le \epsilon\|\mathrm{vec}(A)\|_\infty\right\}. \tag{9.15}$$

Then

$$\kappa(a_k, A) = \|\nabla a_k(A)\|_\infty\|\mathrm{vec}(A)\|_\infty \quad \textit{and} \quad \kappa_{\rm rel}(a_k, A) = \frac{\|\nabla a_k(A)\|_\infty\|\mathrm{vec}(A)\|_\infty}{|a_k(A)|}.$$

The definition of condition number introduced in (9.14) and (9.15) may look non-standard, because of the inclusion of vectorizations. However, the presence of $\mathrm{vec}(E)$ is motivated by (9.1). We have included also $\mathrm{vec}(A)$ in the definition to make it more natural. Moreover, due to the identity

$$\|\mathrm{vec}(M_\sigma)\|_\infty = \|p\|_\infty, \tag{9.16}$$

valid for any Fiedler matrix M_σ, this choice will allow us to get a simpler formula for $\kappa(a_k, M_\sigma)$ (see (9.18) below).

Now, Proposition 9.15, together with (9.7), give us the following formulas for $\kappa(a_k, A)$ and $\kappa_{\rm rel}(a_k, A)$.

Corollary 9.16. *Let $A \in \mathbb{C}^{n\times n}$ and let $\kappa(a_k, A)$ and $\kappa_{\rm rel}(a_k, A)$ be the condition numbers defined in* (9.14) *and* (9.15)*, respectively. Then, for $k = 0, 1, \dots, n-1$,*

$$\kappa(a_k, A) = \|\mathrm{vec}(p_{n-k-1}(A))\|_1\|\mathrm{vec}(A)\|_\infty \quad \textit{and} \quad \kappa_{\rm rel}(a_k, A) = \frac{\|\mathrm{vec}(p_{n-k-1}(A))\|_1\|\mathrm{vec}(A)\|_\infty}{|a_k(A)|}, \tag{9.17}$$

where $p_{n-k-1}(z)$ is the degree $n-k-1$ Horner shift of the polynomial $p(z) := \det(zI - A)$.

Note that, according to (9.17), the relative and absolute condition numbers depend on the norms of A and the degree $n-k-1$ Horner shift evaluated at A of the characteristic polynomial of A. This Horner shift depends in turn on the coefficients $a_{k+1}, \dots, a_{n-1}$ of the characteristic polynomial evaluated at A, namely: $p_{n-k-1}(A) = A^{n-k-1} + a_{n-1}(A)A^{n-k-2} + \cdots + a_{k+1}(A)I$.

In particular, when $A = M_\sigma$ is a Fiedler matrix of a polynomial $p(z)$ as in (1.1), formula (9.17) together with Theorem 9.3 and (9.16), give

$$\kappa(a_k, M_\sigma) = \|p\|_\infty \sum_{i,j=1}^{n} |p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})|, \tag{9.18}$$

where $p_{ij}^{(\sigma,k)}(a_0, a_1, \dots, a_{n-1})$ are given in Theorem 9.3, and they are polynomials of degree at most 2 in the coefficients of p, namely $a_0, \dots, a_{n-1}$.

By considering the maximum condition numbers of all coefficients of the characteristic polynomial we arrive to the following notion.

Definition 9.17. *Let $A \in \mathbb{C}^{n\times n}$ and set $p(z) = \det(zI - A)$. Let $\kappa(a_k, A)$ and $\kappa_{\rm rel}(a_k, A)$ be the condition numbers defined in* (9.14) *and* (9.15)*, respectively. We define the condition number and the relative condition number of the characteristic polynomial of A with respect to perturbations of A as*

$$\kappa(p, A) = \max_{k=0,1,\dots,n-1}\kappa(a_k, A) \quad \textit{and} \quad \kappa_{\rm rel}(p, A) = \max_{k=0,1,\dots,n-1}\kappa_{\rm rel}(a_k, A). \tag{9.19}$$

The following result provides bounds for the absolute and relative condition numbers of the characteristic polynomial when A is a Fiedler matrix.

Proposition 9.18. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\sigma : \{0, 1, \dots, n-1\} \to \{1, \dots, n\}$ be a bijection, let M_σ be a Fiedler companion matrix of $p(z)$ associated with σ, and let $\kappa(p, M_\sigma)$ and $\kappa_{\rm rel}(p, M_\sigma)$ be as in* (9.19)*. Then,*

$$\|p\|_\infty^2 \le \kappa(p, M_\sigma) \le n^3\|p\|_\infty^3 \quad \textit{and} \quad \frac{\|p\|_\infty^2}{\max\{|a_0|, |a_1|, \dots, |a_{n-1}|\}} \le \kappa_{\rm rel}(p, M_\sigma) \le \frac{n^3\|p\|_\infty^3}{\min\{|a_0|, |a_1|, \dots, |a_{n-1}|\}}.$$

Moreover, if $C = C_1, C_2$ denotes both the first and second Frobenius companion matrices, then

$$\|p\|_\infty^2 \leq \kappa(p, C) \leq n^3\|p\|_\infty^2 \quad \textit{and} \quad \frac{\|p\|_\infty^2}{\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}} \leq \kappa_{\rm rel}(p, C) \leq \frac{n^3\|p\|_\infty^2}{\min\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}}.$$

Proof. The bound $\kappa(a_k, M_\sigma) \leq n^3\|p\|_\infty^3$ follows immediately from (9.18) and the bound $|p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})| \leq n\|p\|_\infty^2$ (see Corollary 9.6), valid for all $i, j = 1, \ldots, n$.

From (9.18) and Lemma 9.7, it follows that $\kappa(a_k, M_\sigma) \geq (k+1)|a_{k+1}| \cdot \|p\|_\infty$, for $k = 0, 1, \ldots, n-1$, and $\kappa(a_0, M_\sigma) \geq |a_0| \cdot \|p\|_\infty$. Therefore

$$\kappa(p, M_\sigma) = \max_{k=0,1,\ldots,n-1} \kappa(a_k, M_\sigma) \geq \|p\|_\infty^2.$$

Finally, from

$$\frac{\kappa(a_k, M_\sigma)}{\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}} \leq \frac{\kappa(a_k, M_\sigma)}{|a_k|} \leq \frac{\kappa(a_k, M_\sigma)}{\min\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}}$$

we get the bounds for $\kappa_{\rm rel}(p, M_\sigma)$ in the statement.

For the Frobenius companion matrices, we just note that as a consequence of Corollary 9.6, we have $|p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})| \leq n\|p\|_\infty$, where σ is the permutation corresponding to either the first or the second Frobenius companion matrix. □

Remark 9.19. The factor n^3 appearing in all upper bounds in Proposition 9.18 usually overestimate the condition numbers. It is due to an n^2 factor coming from the maximum possible number of nonzero polynomials $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$ in the sum of the right-hand side in (9.18). This number is usually much less than n^2. For instance, it is equal to $(k+1)(2n-2k-1)$ for the first and second Frobenius companion matrices, as can be seen from (9.10). It is also $(k+1)(2n-2k-1)$ for the coefficients a_k with $k = 2, \ldots, n-1$, equal to $3n-4$ for a_1 and equal to n for a_0, for the Fiedler matrix F in Theorem 9.13, as can be seen by looking at the matrices $p_{n-k-1}(F)$ in Section 9.1.2.

9.2.1 Balancing and condition number

Though similar matrices have the same characteristic polynomial, the sensitivity of its coefficients may be quite different. In other words, the condition numbers $\kappa(a_k, A)$ and $\kappa_{\rm rel}(a_k, A)$ defined in (9.14) and (9.15) are not invariant under diagonal similarity. Since $q(SAS^{-1}) = Sq(A)S^{-1}$, for any polynomial $q(z)$ and any invertible matrix S, formula (9.17) gives

$$\kappa(a_k, SAS^{-1}) = \|\mathrm{vec}(Sp_{n-k-1}(A)S^{-1})\|_1 \|\mathrm{vec}(SAS^{-1})\|_\infty \tag{9.20}$$

and

$$\kappa_{\rm rel}(a_k, SAS^{-1}) = \frac{\|\mathrm{vec}(Sp_{n-k-1}(A)S^{-1})\|_1 \|\mathrm{vec}(SAS^{-1})\|_\infty}{|a_k(A)|}.$$

The norms of the vectors in the right hand side of the previous expression can be quite different for different matrices S. The optimal balancing for a given A (or, equivalently, a given polynomial $p(z) = \det(zI - A)$) from the point of view of the sensitivity of the characteristic polynomial (or, equivalently, from the point of view of backward errors of the root-finding problem via eigenvalue methods) would be given by some nonsingular diagonal matrix D such that $\kappa_{\rm rel}(p, DAD^{-1})$ is minimal among all nonsingular diagonal matrices D (see [130] for the eigenvalue problem). In the particular case of Fiedler matrices, the following result provides a lower bound for this minimal conditioning.

Proposition 9.20. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial of degree n, let $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, let M_σ be the Fiedler matrix of $p(z)$ associated with σ and let $D \in \mathbb{C}^{n\times n}$ be a diagonal and nonsingular matrix. Then, for $k = 0, 1, \ldots, n-1$,*

$$\kappa(a_k, DM_\sigma D^{-1}) \geq (k+1)|a_{n-1}||a_{k+1}| \quad \textit{and} \quad \kappa_{\text{rel}}(a_k, DM_\sigma D^{-1}) \geq \frac{(k+1)|a_{n-1}||a_{k+1}|}{|a_k|},$$

where we set $a_n = 1$.

Proof. We prove the result for $\kappa(a_k, DM_\sigma D^{-1})$, since the bound for the relative condition number can be obtained just dividing by $|a_k|$. The result is a consequence of the fact that diagonal similarity does not change the diagonal entries of a matrix. From (9.20),

$$\begin{aligned}\kappa(a_k, DM_\sigma D^{-1}) \geq & \|\text{diag}(Dp_{n-k-1}(M_\sigma)D^{-1})\|_1 \cdot \|\text{diag}(DM_\sigma D^{-1})\|_\infty \\ = & \|\text{diag}(p_{n-k-1}(M_\sigma))\|_1 \cdot \|\text{diag}(M_\sigma)\|_\infty .\end{aligned}$$

Now we prove that $\text{diag}(M_\sigma) = (-a_{n-1}, 0, \ldots, 0)$ and $\text{diag}(p_{n-k-1}(M_\sigma)) = (0, \ldots, 0, a_{k+1}, \ldots, a_{k+1})$, where the coefficient a_{k+1} appears $(k+1)$ times.

For the diagonal of M_σ the proof proceeds by induction in n. The case $n = 2$ is immediate, since the only possible M_σ are $\begin{bmatrix} -a_1 & -a_0 \\ 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} -a_1 & 1 \\ -a_0 & 0 \end{bmatrix}$. We assume that the identity is true for Fiedler matrices associated with polynomials of degree $n-1$. For degree n, we have to distinguish two cases.

(a) If σ has a consecution at $n-2$ then, using MATLAB notation for columns and rows, M_σ may be written as,

$$M_\sigma = \begin{bmatrix} -a_{n-1} & 1 & 0 \\ W(:,1) & 0 & W(:, 2:n-1) \end{bmatrix},$$

where $W \in \mathbb{C}^{(n-1)\times(n-1)}$ is a Fiedler companion matrix of the polynomial $z^{n-1} + \sum_{k=0}^{n-2} a_k z^k$ (see Theorem 2.16). Therefore, $\text{diag}(M_\sigma) = (-a_{n-1}, 0, W(2,2), W(3,3), \ldots, W(n-1, n-1)) = (-a_{n-1}, 0, \ldots, 0)$, by induction.

(b) If σ has an inversion at $n-2$ then M_σ may be written as

$$M_\sigma = \begin{bmatrix} -a_{n-1} & W(1,:) \\ 1 & 0 \\ 0 & W(2:n-1,:) \end{bmatrix},$$

where $W \in \mathbb{C}^{(n-1)\times(n-1)}$ is a Fiedler companion matrix of the polynomial $z^{n-1} + \sum_{k=0}^{n-2} a_k z^k$ (see Theorem 2.16). Therefore, $\text{diag}(M_\sigma) = (-a_{n-1}, 0, W(2,2), W(3,3), \ldots, W(n-1, n-1)) = (-a_{n-1}, 0, \ldots, 0)$, by induction.

From Lemma 9.7 and equation (9.7), the (i,i) entry of $p_{n-k-1}(M_\sigma)$ is equal to $p_{ii}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = a_{k+1}$, if $n-1 \geq k \geq n-i$ (that is, $i \geq n-k$), and $p_{ii}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1}) = 0$, otherwise. This concludes the proof. □

Note that there exist polynomials $p(z)$ for which the lower bounds in Proposition 9.20 can be as large as desired. In particular, Proposition 9.20 shows that, for large values of $|a_{n-1}|$, the condition number of any coefficient of the characteristic polynomial of any Fiedler matrix will be large, regardless of the diagonal matrix D.

9.3 Backward stability in the case $\|p\|_\infty \leq 1$

Corollary 9.9 indicates that computing the roots of scalar polynomials as the eigenvalues of an arbitrary Fiedler matrix is not backward stable from the polynomial point of view if $\|p\|_\infty$ is large, even if we compute the eigenvalues using a backward stable algorithm. This is revealed by the presence of the factor $\|p\|_\infty$ in (9.4) and $\|p\|_\infty^2$ in (9.5). However, when $\|p\|_\infty$ is moderate, (9.5) guarantees backward stability. This fact is in accordance with results in [160, p. 576], where the authors prove that solving matrix Polynomial Eigenvalue Problems by applying the QZ algorithm to the Frobenius companion pencil is backward stable, provided that the original matrix polynomial has been previously scaled so that all coefficients have norm less than or equal to 1. For scalar polynomials (not necessarily monic), this condition can be always achieved by dividing all coefficients of the original polynomial $p(z)$ by some sufficiently large number. However, if we want to restrict ourselves to the set of monic polynomials to use the QR algorithm, this is not a valid strategy any more, since we could get a non-monic polynomial after dividing the coefficients of $p(z)$ (monic). In order to keep the polynomial $p(z)$ as in (1.1) within the set of monic polynomials, we can consider another kind of scaling as, for instance:

$$\widehat{p}(z) := \alpha^n p(z/\alpha) = z^n + \sum_{k=0}^{n-1} a_k \alpha^{n-k} z^k.$$

Now, α can be chosen so that $|a_k \alpha^{n-k}| \leq 1$, for all $k = 0, 1, \ldots, n-1$. Note that the roots of $p(z)$ can be easily recovered from those of $\widehat{p}(z)$ just dividing by α. Once all coefficients of $\widehat{p}(z)$ have absolute value less than or equal to 1, we can apply the QR algorithm to any Fiedler companion matrix of $\widehat{p}(z)$ to get its roots, and then recover the roots of $p(z)$. However, this does not guarantee that the method is backward stable. It is not difficult to find examples of quadratic polynomials $p(z)$ such that there is a polynomial $\widehat{q}(z)$ with $\|\widehat{p} - \widehat{q}\| = O(u)\|\widehat{p}\|$, but $\|p - q\|/\|p\|$ is $O(1)$, with $q(z) = (1/\alpha^2)\widehat{q}(\alpha z)$.

We want to emphasize that we are not considering in this chapter the backward errors of single roots of p, but the backward error of the set of all roots of p. Backward errors of single roots has been considered in [147] for the more general case of matrix Polynomial Eigenvalue Problems. In particular, the backward error of a single computed root $\widetilde{\lambda}$ considered in [147] is:

$$\eta(\widetilde{\lambda}) = \min\left\{\epsilon : \ (p + \Delta p)(\widetilde{\lambda}) = 0, \quad |\Delta a_i| \leq \epsilon |a_i|, \ i = 0, 1, \ldots, n\right\},$$

where $p(z) = \sum_{k=0}^n a_k z^k$, and $\Delta p(z) = \sum_{k=0}^n (\Delta a_k) z^k$ are not necessarily monic. It is shown in [147, Theorem 7] that, for quadratic matrix polynomials all whose coefficients have 2-norm equal to 1, computing the eigenvalues of its companion pencil (defined in [147, p. 347]) with a backward stable eigenvalue algorithm gives, from the polynomial point of view, a coefficientwise backward stable method for the Quadratic Eigenvalue Problem. Though, as we have mentioned above, we are considering different notions of backward error, this fact seems to be in accordance with Corollary 9.9 when $\|p\|_\infty = 1$ and with the discussion right below.

We also emphasize that the backward stability of polynomial root-finding when $\|p\|_\infty = 1$ does not guarantee small relative backward errors in each coefficient. In other words, we can not guarantee that

$$\max_{k=0,1,\ldots,n-1} \frac{|\widetilde{a}_k - a_k|}{|a_k|} = O(u) \tag{9.21}$$

even in the case $\|p\|_\infty = 1$. In Section 9.4 we show some numerical experiments where $\|p\|_\infty = 1$ and (9.21) does not hold. However, when $|a_k|$ is moderate, for all $k = 0, 1, \ldots, n-1$, and not too close to zero (loosely speaking, of order $\Theta(1)$), then (9.4)–(9.5) imply that (9.21) holds, also in accordance with [147].

9.4 Numerical experiments

In this section we provide numerical experiments that support our theoretical results. In particular, our goals are: (i) to show whether or not the bounds in (9.4)–(9.5) correctly predict the dependence on the norm of $p(z)$ of the largest backward error that may be obtained if the roots of $p(z)$ are computed as the eigenvalues of a Fielder matrix with a backward stable eigenvalue algorithm; (ii) to show that if the roots of a polynomial $p(z)$, with moderate coefficients, are computed as the eigenvalues of a Fiedler matrix, then this process is normwise backward stable, regardless of the Fiedler matrix that is used, which implies that, in this situation, any Fiedler matrix can be used for the root-finding problem with the same reliability as the Frobenius companion matrices; (iii) to investigate, from the point of view of backward errors, the effect of balancing Fiedler matrices; and (iv) following [53], to show that Theorem 9.3 may be used to predict the backward error when the roots of a monic polynomial are computed as the eigenvalues of a Fiedler matrix. Along this section we denote by $u = 2^{-52}$ the machine epsilon in IEEE double precision arithmetic.

Given a monic polynomial $p(z)$ of degree n, we denote by $\{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_n\}$ the roots of $p(z)$ computed as eigenvalues of a Fiedler matrix M_σ using a backward stable eigenvalue algorithm. In our case, the eigenvalue algorithm will be the QR eigenvalue algorithm as implemented in the command `eig` of MATLAB. If we denote by $\widetilde{p}(z)$ the monic polynomial of degree n whose roots are $\{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_n\}$, namely, $\widetilde{p}(z) = \prod_{k=1}^{n}(z - \widetilde{\lambda}_k) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k$, then we are interested in the following quantities:

- the normwise backward error (NBE): $\|\widetilde{p} - p\|_\infty / \|p\|_\infty$, and
- the coefficientwise backward error (CBE): $\max_{k=0,1,\ldots,n-1} (|\widetilde{a}_k - a_k|/|a_k|)$.

In the numerical experiments, we consider monic polynomials of degree 20 and the following Fiedler companion matrices associated with degree-20 polynomials:

(a) the second Frobenius companion matrix $C_2 = M_{\sigma_1}$ with $\mathrm{PCIS}(\sigma_1) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$,

(b) the pentadiagonal Fiedler matrix $P_1 = M_{\sigma_2}$ with $\mathrm{PCIS}(\sigma_2) = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$,

(c) the Fiedler matrix $F = M_{\sigma_3}$ with $\mathrm{PCIS}(\sigma_3) = (0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$, and

(d) the Fiedler matrix M_{σ_4} with $\mathrm{PCIS}(\sigma_4) = (1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1)$.

Recall that the matrices M_{σ_2} and M_{σ_3} are the Fiedler matrices considered in (2.7) and (2.8), respectively.

Given a monic polynomial $p(z)$ of degree 20 and a Fiedler matrix M_σ associated with $p(z)$, to compute the polynomial $\widetilde{p}(z)$ we proceed as follows. First, we compute the eigenvalues of M_σ using the function `eig` in MATLAB (with and/or without balancing, see comments below); then, if $\{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_{20}\}$ denote the computed eigenvalues, we compute the polynomial $\widetilde{p}(z) = \prod_{k=1}^{20}(z - \widetilde{\lambda}_k) = z^{20} + \sum_{k=0}^{19} \widetilde{a}_k z^k$ using the function `vpa` (variable precision arithmetic) followed by the command `poly` on a diagonal matrix whose diagonal entries are $\{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_{20}\}$, in MATLAB with 32 decimal digits of accuracy.

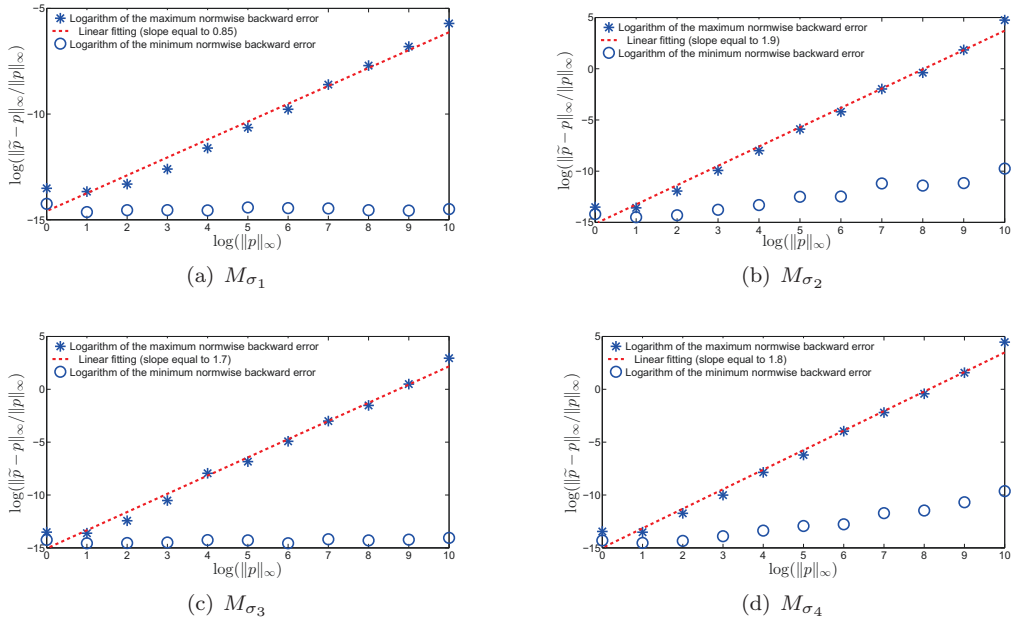
9.4.1 Numerical experiments that show the dependence of the normwise backward error with $\|p\|_\infty$

In this subsection, we perform numerical experiments to determine whether or not the largest normwise backward errors that may be obtained if the roots of monic polynomials are computed

as the eigenvalues of a Fiedler matrix M_σ with the command `eig`, behave like $\|\widetilde{p}-p\|_\infty/\|p\|_\infty = O(u)\|p\|_\infty^2$, when M_σ is a Fiedler matrix other than the Frobenius ones, or like $\|\widetilde{p}-p\|_\infty/\|p\|_\infty = O(u)\|p\|_\infty$, when M_σ is one of the Frobenius companion matrices, as it is predicted by Corollary 9.9. We perform numerical experiments with and without balancing the Fiedler matrices. Our results show that if we do not balance the Fiedler matrices the bound in Corollary 9.9, although in a lot of cases is very pessimistic, predicts well the dependence with $\|p\|_\infty$ of the largest backward errors. If the Fiedler matrices are balanced, our results show that there is still a dependence with $\|p\|_\infty$ of the largest normwise backward errors, and that this dependence is similar for all Fiedler matrices. Also we show that the backward errors that are usually obtained when the Fiedler matrices are balanced are almost independent of the norm of the polynomials, and that polynomial root-finding algorithms using balanced Fiedler matrices are usually normwise backward stable.

In order to see the dependence of the backward error with $\|p\|_\infty$ we proceed as follows. For each $k = 0, 1, \ldots, 10$ we generate 500 random degree-20 polynomials with coefficients of the form $a \cdot 10^c$, where a is drawn from the uniform distribution on the interval $[-1, 1]$ and c is drawn from the uniform distribution on the interval $[-k, k]$, also we set $a_0 = 10^k$. The reasons to set $a_0 = 10^k$ is to fix the infinity norm of the 500 random polynomials to be 10^k. For each of these 11 samples of 500 random polynomials, we compute the normwise backward errors, as it is explained at the beginning of Section 9.4, when their roots are computed as the eigenvalues of the four Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, with and without balancing them.

In Figures 9.4.1-(a), 9.4.1-(b), 9.4.1-(c), and 9.4.1-(d) we plot the decimal logarithms of the maximum and the minimum normwise backward errors obtained for each of the 11 samples of 500 random polynomials against the logarithms of the norm of the polynomials, when their roots are computed as the eigenvalues of $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, respectively, without balancing them. We also plot a linear fitting for the logarithms of the maximum normwise backward errors in order to get the dependence with $\|p\|_\infty$.



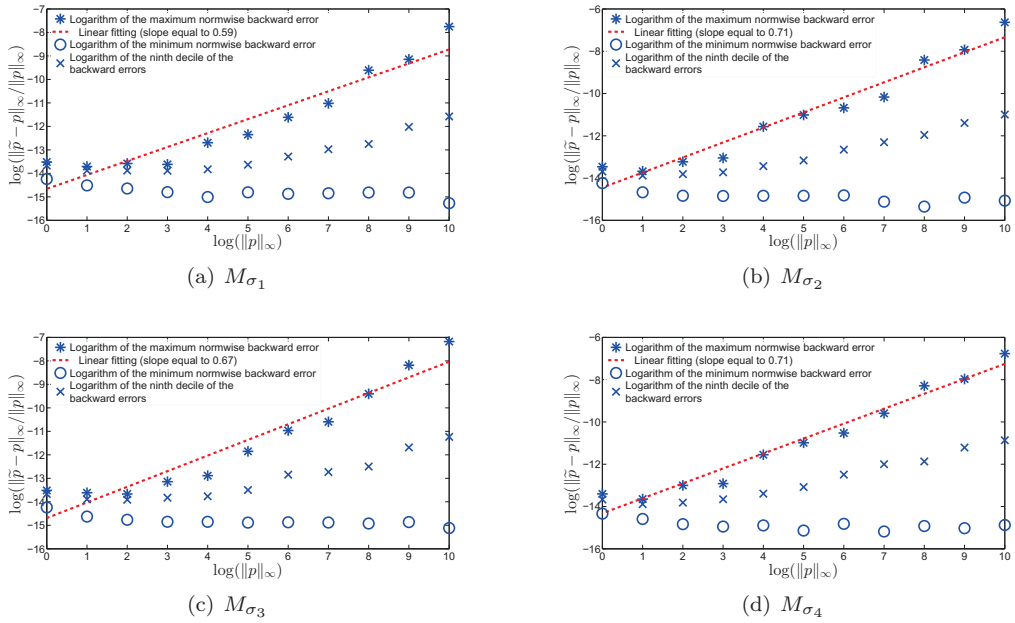
(a) M_{σ_1} (b) M_{σ_2}

(c) M_{σ_3} (d) M_{σ_4}

Figure 9.4.1: Decimal logarithms of the maximum and minimum normwise backward errors obtained for each of the 11 samples of 500 random degree-20 polynomials, for $k = 0, 1, \ldots, 10$, with a fixed infinite norm equal to 10^k and with coefficients of the form $a \cdot 10^c$, where a is drawn from the uniform distribution on $[-1, 1]$ and c is drawn from the uniform distribution on $[-k, k]$, and where we set $a_0 = 10^k$, when their roots are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing them.

As may be seen in Figures 9.4.1-(a), 9.4.1-(b), 9.4.1-(c), and 9.4.1-(d), there is a dependence with $\|p\|_\infty$ of the largest normwise backward errors of the form $\|p\|_\infty^\alpha$. From the linear fittings we obtain $\alpha = 0.85$ for $M_{\sigma_1} = C_2$, $\alpha = 1.9$ for $M_{\sigma_2} = P_1$, $\alpha = 1.7$ for $M_{\sigma_3} = F$, and $\alpha = 1.8$ for M_{σ_4}. This is consistent with the bound in Corollary 9.9, which predicts $\alpha = 1$ for the Frobenius companion matrices C_1 and C_2, and $\alpha = 2$ for Fiedler matrices other than the Frobenius ones. Also note that in Figures 9.4.1-(a),9.4.1-(b),9.4.1-(c) and 9.4.1-(d) it may be seen that the bound in Corollary 9.9 is in some cases very pessimistic, since there are polynomials for which we get small normwise backward errors, regardless of their norms.

Next, we investigate the effect of balancing the Fiedler matrices in the backward errors. In Figures 9.4.2-(a), 9.4.2-(b), 9.4.2-(c), and 9.4.2-(d), we plot the decimal logarithms of the maximum and the minimum normwise backward errors obtained for each of the 11 samples of 500 random polynomials against the logarithms of the norm of the polynomials, when their roots are computed as the eigenvalues of $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, respectively, but in this case the Fiedler matrices are balanced before we compute their eigenvalues. As in the previous experiment, we plot a linear fitting for the logarithms of the maximum normwise backward errors in order to get the dependence with $\|p\|_\infty$. We also plot the ninth decile of the normwise backward error for each of the 11 samples.



(a) M_{σ_1} (b) M_{σ_2}

(c) M_{σ_3} (d) M_{σ_4}

Figure 9.4.2: Decimal logarithms of the maximum and minimum normwise backward errors obtained for the 11 samples of 500 random degree-20 polynomials with, for $k = 0, 1, \ldots, 10$, a fixed infinite norm equal to 10^k and with coefficients of the form $a \cdot 10^c$, where a is drawn from the uniform distribution on $[-1, 1]$ and c is drawn from the uniform distribution on $[-k, k]$, and where we set $a_0 = 10^k$, when their roots are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, balancing them before computing their eigenvalues.

As may be seen in Figures 9.4.2-(a), 9.4.2-(b), 9.4.2-(c), and 9.4.2-(d), there is a dependence of the largest backward errors with the norm of the polynomials of the form $\|p\|_\infty^\alpha$, but this dependence is more or less similar for all four Fiedler matrices. In particular, from the linear fittings, we get $\alpha = 0.59$ for $M_{\sigma_1} = C_2$, $\alpha = 0.71$ for $M_{\sigma_2} = P_1$, $\alpha = 0.67$ for $M_{\sigma_3} = F$, and $\alpha = 0.71$ for M_{σ_4}. Also notice that 90% of the backward errors obtained when the roots of the polynomials are computed as the roots of $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$ are excellent, since they are more or less between 10^{-12} and 10^{-16}, even for polynomials with norms as large as 10^{10}.

9.4.2 Numerical experiments with polynomials of moderate coefficients

In this subsection we show that, from the point of view of backward errors, when the coefficients of $p(z)$ are bounded in absolute value by a moderate number, any Fiedler matrix may be used for the root-finding problem with the same reliability as the Frobenius companion matrices. In particular, we provide numerical evidence that supports what we claim in Section 9.3, namely, that computing the roots of a monic polynomial $p(z)$ as in (1.1), with $|a_i|$ moderate, for $i = 0, 1, \ldots, n-1$, as the eigenvalues of a Fiedler matrix using a backward stable eigenvalue algorithm is normwise backward stable, regardless of the Fiedler matrix that is used. In addition, we show that to have $|a_i|$ moderate, for $i = 0, 1, \ldots, n-1$, it is not enough to guarantee coefficientwise backward stability. Finally, we provide numerical evidence that supports the last sentence in Section 9.3, namely, that (9.21) holds when $|a_i| = \Theta(1)$, for $i = 0, 1, \ldots, n-1$, regardless of the Fiedler matrix that is used.

In the first set of numerical experiments, we consider a random sample of 1000 degree-20 polynomials with coefficients drawn from the uniform distribution on the interval $[-100, 100]$, but we set $a_{19} = 10^{-10}$. The reason for setting $a_{19} = 10^{-10}$ is to show that we may have a small normwise backward error but, at the same time, we may have a big coefficientwise backward error. In Table 9.4.1, we give the mean, the maximum and the minimum of the decimal logarithms of the normwise and coefficientwise backward errors (Log-Mean NBE, Log-Maximum NBE, Log-Minimum NBE, Log-Mean CBE, Log-Maximum CBE and Log-Minimum CBE, respectively) obtained when the roots of the polynomials are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing them.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean NBE	-13.6	-12.6	-13.6	-12.5
Log-Maximum NBE	-12.9	-11.7	-12.1	-11.6
Log-Minimum NBE	-14.5	-13.4	-14.3	-13.4
Log-Mean CBE	-3.5	-3.8	-3.6	-3.8
Log-Maximum CBE	-2.7	-2.9	-2.7	-2.9
Log-Minimum CBE	-6.7	-6.9	-7.1	-6.3

Table 9.4.1: Mean, maximum and minimum of the decimal logarithms of the normwise (NBE) and coefficientwise (CBE) backward errors obtained for 1000 random degree-20 polynomials, with coefficients drawn from the uniform distribution on $[-100, 100]$ and setting $a_{19} = 10^{-10}$, when their roots are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing them.

As may be seen in Table 9.4.1, the normwise backward errors obtained when the roots of the polynomials are computed as the eigenvalues of $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$ are excellent for all four Fiedler matrices. But also note that this is not true for the coefficientwise backward errors. These results are consistent with the claims in Section 9.3.

Next, we consider a random sample of 1000 degree-20 polynomials with coefficients of the form 10^{c_1} where c_1 is drawn from the uniform distribution on the interval $[-2, 2]$. In Table 9.4.2 we give the mean, the maximum and the minimum of the decimal logarithms of the normwise and coefficientwise backward errors (Log-Mean NBE, Log-Maximum NBE, Log-Minimum NBE, Log-Mean CBE, Log-Maximum CBE and Log-Minimum CBE, respectively) obtained when the roots of the polynomials are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing them.

As may be seen in Table 9.4.2, the normwise backward errors obtained when the roots of the polynomials are computed as the eigenvalues of $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$ are excellent, as in Table 9.4.1. The coefficientwise backward errors are not so small as the normwise ones, but they are still excellent since we are dealing with polynomials whose coefficients may have absolute values that

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean NBE	-14.1	-13.2	-14.1	-13.3
Log-Maximum NBE	-13.4	-11.8	-12.5	-11.7
Log-Minimum NBE	-14.7	-14.5	-14.8	-14.8
Log-Mean CBE	-11.0	-10.2	-11.0	-10.2
Log-Maximum CBE	-10.0	-8.3	-9.1	-8.4
Log-Minimum CBE	-12.4	-12.2	-12.6	-12.7

Table 9.4.2: Mean, maximum and minimum of the decimal logarithms of the normwise (NBE) and coefficientwise (CBE) backward errors obtained for 1000 random degree-20 polynomials, with coefficients of the form 10^{c_1}, where c_1 is drawn from the uniform distribution on $[-2, 2]$, when their roots are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing them.

differ in four orders of magnitude.

9.4.3 Numerical experiments balancing Fiedler matrices

In this subsection we perform numerical experiments to study, from the point of view of backward errors, the effect of balancing Fiedler matrices. We show that, when a Fiedler matrix M_σ is balanced before computing its eigenvalues, the backward error obtained if we compute the roots of $p(z)$ as the eigenvalues of M_σ may be much smaller than the backward error that is obtained when M_σ is not balanced, regardless of the Fiedler matrix that is used. We show also that balancing a Fiedler matrix is usually enough to guarantee that the process of computing the roots of a polynomial as the eigenvalues of a Fiedler matrix is normwise backward stable, even if the polynomial has large coefficients. Finally, we investigate the effect of the size of the coefficient a_{n-1}, since Proposition 9.20 suggests that it plays a key role in getting or not backward stability after balancing Fiedler matrices. To be precise, Proposition 9.20 shows that, for large values of $|a_{n-1}|$, the condition number of any coefficient of the characteristic polynomial of any Fiedler matrix will be large, regardless of the balancing. This leads us to expect large backward errors when $|a_{n-1}|$ is large.

We consider a random sample of 1000 degree-20 polynomials with coefficients of the form

$$a_1 \cdot 10^{c_1} + i\, a_2 \cdot 10^{c_2}, \tag{9.22}$$

where i denotes the imaginary unit, and a_1, a_2 are drawn from the uniform distribution on the interval $[-1, 1]$ and c_1 and c_2 are drawn from the uniform distribution on the interval $[-10, 10]$. These polynomials, considered in [150], allow us to measure the normwise backward errors with varying orders of magnitude in the coefficients of $p(z)$. We also consider a second sample of 1000 degree-20 polynomials with coefficients of the form (9.22), but we fix $a_{19} = 1$. The reason for considering this second sample is to study the effect of balancing Fiedler matrices when $|a_{n-1}|$ is moderate.

For the first sample of random polynomials, in Tables 9.4.3-(a) and 9.4.3-(b) we give the mean, the maximum and the minimum of the decimal logarithms of the normwise backward errors (Log-Mean NBE, Log-Maximum NBE, Log-Minimum NBE, respectively) obtained when the roots of the polynomials are computed as the eigenvalues of $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, when these Fiedler matrices are not or are balanced, respectively.

Several observations may be drawn from the data in Tables 9.4.3-(a) and 9.4.3-(b). First note, from the data in Log-Maximum NBE in Table 9.4.3-(a), that if the Fiedler matrices are not balanced, the backward errors obtained may be very large. Note also that the largest of these backward errors is consistent with (9.4) for the Frobenius companion matrices, and with (9.5) for Fiedler matrices other than the Frobenius ones. Second, note that the process of balancing the

(a) The Fiedler matrices are not balanced.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean NBE	-10.5	-2.4	-9.9	-3.0
Log-Maximum NBE	-5.8	3.2	0.1	3.5
Log-Minimum NBE	-14.7	-8.9	-14.7	-10.0

(b) The Fiedler matrices are balanced.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean NBE	-13.1	-13.1	-13.1	-12.9
Log-Maximum NBE	-8.1	-7.5	-8.0	-7.8
Log-Minimum NBE	-14.7	-14.9	-15.1	-14.8

Table 9.4.3: Mean, maximum, and minimum of the decimal logarithms of the normwise backward errors obtained for a sample of 1000 random degree-20 polynomials, with coefficients of the form (9.22), when their roots are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing and with balancing.

Fiedler matrices makes that the backward errors obtained after balancing may be much smaller than the backward errors obtained when the Fiedler matrices are not balanced (this is especially evident for M_{σ_2} and M_{σ_3}). Finally, note, from the data in Log-Maximum NBE in Table 9.4.3-(b), that there are polynomials for which balancing the Fiedler matrices does not guarantee that the process of computing their roots as the eigenvalues of Fiedler matrices is normwise backward stable.

In Tables 9.4.4-(a) and 9.4.4-(b) we display the mean, the maximum and the minimum of the decimal logarithms of the normwise backward errors (Log-Mean NBE, Log-Maximum NBE, Log-Minimum NBE, respectively) that are obtained when the roots of the polynomials of the second sample are computed as the eigenvalues of the four Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, when the Fiedler matrices are not or are balanced, respectively. Recall that for this sample of degree-20 random polynomials we set $a_{19} = 1$.

(a) The Fiedler matrices are not balanced.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean NBE	-6.9	-3.2	-6.9	-3.4
Log-Maximum NBE	-5.6	3.0	-3.4	3.0
Log-Minimum NBE	-9.8	-10.6	-9.9	-11.1

(b) The Fiedler matrices are balanced.

	M_{σ_1}	M_{σ_2}	M_{σ_3}	M_{σ_4}
Log-Mean NBE	-13.9	-13.9	-13.9	-13.7
Log-Maximum NBE	-11.6	-11.1	-11.6	-10.4
Log-Minimum NBE	-15.1	-14.8	-15.0	-15.0

Table 9.4.4: Mean, maximum, and minimum of the decimal logarithms of the normwise backward errors obtained when the roots of the polynomials of the second sample of random polynomials (i.e., coefficients from (9.22) and $a_{19} = 1$) are computed as the eigenvalues of the four Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, without balancing and with balancing.

As in the first sample of random polynomials, we may see in Tables 9.4.4-(a) and 9.4.4-(b) that the backward errors obtained when the Fiedler matrices are not balanced may be very large. Also,

we may see that the backward errors may be much smaller when the Fiedler matrices are balanced. Finally note that for this second sample the largest backward errors obtained when the Fiedler matrices are balanced are smaller than the largest ones obtained for the first sample.

9.4.4 Using Theorem 9.3 to predict the coefficientwise backward error

In this subsection we show that Theorem 9.3 can be used to predict the coefficientwise backward error, without computing explicitly the polynomial $\widetilde{p}(z)$ (something that may not be possible for high degree polynomials, since using `vpa` makes this process very slow), and that this backward error is usually small for all Fiedler matrices if the process of balancing is used. Of course, the normwise backward error can be also predicted from Theorem 9.3, but we omit it for brevity. As in Section 8.6.3, we explore the following degree-20 monic polynomials:

($p1$) the Wilkinson polynomial: $p(z) = \prod_{k=1}^{20}(z-k)$,

($p2$) the monic polynomial with zeros: $-2, -1.8, -1.6, \ldots, 1.6, 1.8$,

($p3$) $p(z) = (20!)\sum_{k=0}^{20} z^k/k!$,

($p4$) the Bernoulli polynomial of degree 20,

($p5$) $p(z) = \sum_{k=0}^{20} z^k$,

($p6$) the monic polynomial with zeros $2^{-10}, 2^{-9}, \ldots, 2^8, 2^9$,

($p7$) the Chebyshev polynomial of degree 20,

($p8$) the monic polynomial with zeros equally spaced on a sine curve, that is,

$$p(z) = \prod_{k=-10}^{9} \left(z - \frac{2\pi}{19}(k+0.5) - i \cdot \sin \frac{2\pi}{19}(k+0.5) \right).$$

Also, we consider again the four Fiedler companion matrices associated with degree-20 polynomials introduced at the beginning of Section 9.4, namely $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$.

We repeat the numerical experiments performed in [53]. Our results show that Theorem 9.3 always predicts a small componentwise backward error, regardless of the Fiedler matrix that is used, and that this predicted backward error is usually pessimistic by at most one, two or three orders of magnitude, except for the polynomial $p6$, where the predicted backward error is pessimistic by 6 orders of magnitude. Note that in this case the ratio $(|a_{19}| \cdot |a_1|)/|a_0|$ is of order 2^{19}, so Proposition 9.20 ensures that the condition number for the coefficient a_0 is large. However, the perturbations in the numerical experiments does not seem to affect this coefficient in such a severe way.

In order to use Theorem 9.3 to predict the coefficientwise backward error, we need to model the backward error introduced by the algorithm for computing the eigenvalues of a Fiedler matrix. Since standard eigenvalue algorithms first balance the matrix, if we set $B_\sigma := DM_\sigma D^{-1}$, where D is the diagonal matrix that balances M_σ, then a backward stable eigenvalue algorithm applied to a Fiedler matrix M_σ computes the exact eigenvalues of the matrix $B_\sigma + \widetilde{E}$, with $\|\widetilde{E}\| = O(u)\|B_\sigma\|$. Due to these considerations, we model the backward error introduced by a backward stable eigenvalue algorithm applied to M_σ by means of an error matrix $\widetilde{E} = (\widetilde{E}_{ij})$, with

$$\widetilde{E}_{ij} = 2^{-52} \cdot \|B_\sigma\|_2 \cdot \epsilon_{ij} \quad \text{for } i,j = 1,2,\ldots,20, \tag{9.23}$$

where ϵ_{ij} is drawn from the uniform distribution on the interval $[-1,1]$. If we denote by $\{\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_n\}$ the eigenvalues of $B_\sigma + \widetilde{E}$ then, since a similarity transformation does not change the characteristic polynomial of a matrix, these eigenvalues are the roots of the characteristic polynomial

of $D^{-1}(B_\sigma + \widetilde{E})D = M_\sigma + E$, where $E = D^{-1}\widetilde{E}D$, that is, they are the roots of the polynomial $\det(zI - M_\sigma - E) = \widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_n z^k$. Then, we can use Theorem 9.3 to compute the coefficients of $\widetilde{p}(z)$, up to first order in E, and use $\max_{a_k \neq 0} |\widetilde{a}_k - a_k|/|a_k|$ as a prediction of the coefficientwise backward error. Finally we can compare this predicted backward error with the observed one, computed as explained at the beginning of Section 9.4.

In Table 9.4.5, we display the decimal logarithms of the predicted and the observed coefficientwise backward error (Log Predicted CBE and Log Observed CBE, respectively), when the roots of the polynomials $p1$-$p8$ are computed as the eigenvalues of the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$.

(a) M_{σ_1}

	$p1$	$p2$	$p3$	$p4$	$p5$	$p6$	$p7$	$p8$
Log Predicted CBE	-12.0	-10.4	-13.4	-13.3	-13.7	-8.8	-12.3	-13.6
Log Observed CBE	-13.9	-13.5	-14.2	-13.8	-13.9	-13.8	-14.7	-14.6

(b) M_{σ_2}

	$p1$	$p2$	$p3$	$p4$	$p5$	$p6$	$p7$	$p8$
Log Predicted CBE	-12.3	-12.1	-10.3	-13.3	-13.4	-9.3	-13.2	-13.3
Log Observed CBE	-13.8	-14.0	-12.0	-13.8	-13.6	-14.1	-13.7	-13.9

(c) M_{σ_3}

	$p1$	$p2$	$p3$	$p4$	$p5$	$p6$	$p7$	$p8$
Log Predicted CBE	-12.1	-12.9	-13.5	-13.0	-13.7	-8.8	-12.3	-13.5
Log Observed CBE	-14.0	-13.8	-13.7	-14.1	-13.9	-13.8	-14.7	-14.3

(d) M_{σ_4}

	$p1$	$p2$	$p3$	$p4$	$p5$	$p6$	$p7$	$p8$
Log Predicted CBE	-12.3	-12.8	-12.8	-13.5	-13.3	-9.6	-13.9	-13.7
Log Observed CBE	-14.0	-14.2	-13.4	-14.1	-13.9	-14.0	-15.1	-14.1

Table 9.4.5: Decimal logarithms of the predicted and observed coefficientwise backward error for the Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$ of the eight polynomials $p1$-$p8$.

As may be seen in Table 9.4.5, the coefficientwise backward errors are well predicted by Theorem 9.3, with the exception of the polynomial $p6$. In [53] it was also observed that the coefficientwise backward error for $p6$, when the Frobenius companion matrix is used to compute its roots, was far more favorable than the predicted one. For this polynomial and for the four Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}, M_{\sigma_4}$, the coefficientwise backward error comes from $|\widetilde{a}_0 - a_0|/|a_0|$. The most important conclusion to be extracted from Table 9.4.5 for our purposes is that the four Fiedler matrices $M_{\sigma_1}, M_{\sigma_2}, M_{\sigma_3}$, and M_{σ_4} behave equally well from the point of view of backward errors in polynomials $p1 - p8$.

9.5 The Sylvester space of a Fiedler matrix

The study of the geometry of matrix spaces sheds light on the explanation of numerical processes involving matrices or matrix pencils. In particular, the theory of orbits has been used in the analysis of errors of the algorithms for computing eigenvalues and canonical forms (see [7], [54, 55] and [53]). In this section, and inspired by the motivating paper [53], we analyze from a geometrical

point of view the polynomial root-finding problem solved as an eigenvalue problem with Fiedler companion matrices. Our main result is Theorem 9.24, where we prove that the space of Sylvester matrices associated with a given Fiedler matrix M_σ is transversal to the similarity orbit of M_σ. This result extends the corresponding one for Frobenius companion matrices [53, Prop. 2.1].

Let $p(z)$ be a monic polynomial as in (1.1) and let M_σ be a Fiedler matrix of $p(z)$. Let us consider the Euclidean matrix space $\mathbb{C}^{n\times n}$ with the usual Frobenius inner product

$$(A, B) = \mathrm{tr}(AB^*),$$

where M^* denotes the conjugate transpose of $M \in \mathbb{C}^{n\times n}$. In this space, the set of matrices similar to a given matrix $A \in \mathbb{C}^{n\times n}$ is a differentiable manifold in $\mathbb{C}^{n\times n}$. This manifold is the orbit of A under the action of similarity:

$$\mathcal{O}(A) := \{SAS^{-1} : \det(S) \neq 0\}.$$

We will refer to the elements of a manifold as *points*, even though all manifolds considered in this section are manifolds whose points are matrices.

It is known that the tangent space of $\mathcal{O}(A)$ at A is the set

$$T_A\mathcal{O}(A) := \{AX - XA \text{ for some } X \in \mathbb{C}^{n\times n}\}.$$

The *normal space* of $\mathcal{O}(A)$ at A, denoted by $N_A\mathcal{O}(A)$, is the set of matrices orthogonal to any matrix in $T_A\mathcal{O}(A)$:

$$N_A\mathcal{O}(A) := \{Y \in \mathbb{C}^{n\times n} \text{ such that } (Y, V) = 0, \text{ for all } V \in T_A\mathcal{O}(A)\},$$

and the *centralizer of* A is the set of matrices commuting with A:

$$C(A) := \{X \in \mathbb{C}^{n\times n} \text{ such that } AX - XA = 0\}$$

The following facts are already known:

(a) $C(A^*) = N_A\mathcal{O}(A)$ (see [7, Lemma, p. 34]).

(b) If A is a non-derogatory matrix, then:

(b1) $C(A) = \{q(A) : q \text{ is a polynomial}\}$ (see [87, Th. 3.2.4.2]).

(b2) $\dim C(A) = n$ (see [7, Corollary, p. 35]).

(c) M_σ is a non-derogatory matrix, for all σ.

For claim (c), just recall that M_σ is similar to C_1, and that C_1 is non-derogatory (see [87, p. 147]).

As a consequence of claims (a)–(c) above, we have that $\dim N_{M_\sigma}\mathcal{O}(M_\sigma) = n$, for all σ, so there is a basis of $N_{M_\sigma}\mathcal{O}(M_\sigma)$ consisting of n matrices which are polynomials in M_σ^*. In Proposition 9.21, we state that.

Proposition 9.21. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial, $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, M_σ be the Fiedler matrix of $p(z)$ associated with the bijection σ, and let $p_d(z)$ be the dth Horner shift of $p(z)$, for $d = 0, 1, \ldots, n-1$. Set $p_0(M_\sigma) = I_n$ and*

$$p_{n-k}(M_\sigma) = M_\sigma^{n-k} + a_{n-1}M_\sigma^{n-k-1} + \cdots + a_{k+1}M_\sigma + a_k I, \quad \text{for } k = 1, \ldots, n-1.$$

Then $\{p_k(M_\sigma)^*\}_{k=0}^{n-1}$ is a basis for $N_{M_\sigma}\mathcal{O}(M_\sigma)$.

Note that the set $\{p_k(M_\sigma)^*\}_{k=0}^{n-1}$ is linearly independent because, since M_σ is non-derogatory, its minimal polynomial coincides with its characteristic polynomial. Any n linearly independent polynomials in M_σ^* would serve as a basis for $N_{M_\sigma}\mathcal{O}(M_\sigma)$, but in Section 9.1.1 we have seen that the matrices $p_k(M_\sigma)$ play an important role in determining how the coefficients of the characteristic polynomial of M_σ change when the matrix is perturbed (see (9.7)).

First order perturbations of the coefficients of $p(z)$, with $p(z) = \det(zI - C_1)$, have been studied in [53]. To do so, the authors decompose the perturbation matrix E as

$$E = E^{\text{tan}} + E^{\text{syl}}, \tag{9.24}$$

where E^{tan} belongs to the tangent space to $\mathcal{O}(C_1)$ at C_1 and E^{syl} is of the form

$$E^{\text{syl}} = \begin{bmatrix} E_{11} & \dots & E_{1n} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

The matrix E^{syl} belongs to the tangent space (at any point) to the *Sylvester space* of C_1. We recall that the (affine) Sylvester space of C_1 is the set of all matrices of the form

$$\begin{bmatrix} E_{11} & E_{12} & \dots & E_{1n} \\ 1 & 0 & \dots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{bmatrix},$$

that is, the set of "all first Frobenius companion matrices"[2]. It may be proved that, to first order in E, the matrix E^{tan} does not affect the coefficients of $p(z)$. Below, we prove an equivalent result for any Fiedler matrix M_σ. For this, we first define the Sylvester space of any Fiedler matrix, which is a natural generalization of the Sylvester space of C_1.

Definition 9.22. (Sylvester space of a Fiedler matrix) *Let $\sigma : \{0, 1, \dots, n-1\} \to \{1, \dots, n\}$ be a bijection. Then, the (affine) Sylvester space associated with the bijection σ, denoted by* $\text{Syl}(\sigma)$*, is the set of Fiedler matrices associated with σ, that is,*

$$\text{Syl}(\sigma) := \left\{ M_\sigma(p) : \; p(z) = z^n + \sum_{k=0}^{n-1} c_k z^k, \quad c_k \in \mathbb{C} \right\},$$

where $M_\sigma(p)$ is the matrix in (2.1)*.*

For example, the Sylvester space associated with the bijection σ, such that $\text{PCIS}(\sigma) = (1, 1, 1, 0, 0, 0)$, is the set of matrices of the form

$$\begin{bmatrix} c_6 & c_5 & c_4 & c_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_2 & 0 & 1 & 0 \\ 0 & 0 & 0 & c_1 & 0 & 0 & 1 \\ 0 & 0 & 0 & c_0 & 0 & 0 & 0 \end{bmatrix},$$

[2]We note that the companion matrix considered in [53] is not exactly C_1, but the companion matrix obtained from C_1 after performing a symmetry through the main anti-diagonal, and accordingly with the Sylvester space.

where $c_k \in \mathbb{C}$, for $k = 0, 1, \ldots, 6$, may take any value. The tangent space of $\mathrm{Syl}(\sigma)$ at a given point, denoted by $\mathrm{TSyl}(\sigma)$, is the set of matrices that we get if we remove the entries identically equal to 1 in the matrix above. In other words, the underlying vector space to the affine space. For example, for the previous bijection σ, the tangent space of $\mathrm{Syl}(\sigma)$ is the set of matrices of the form

$$\begin{bmatrix} c_6 & c_5 & c_4 & c_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_0 & 0 & 0 & 0 \end{bmatrix},$$

where $c_k \in \mathbb{C}$, for $k = 0, 1, \ldots, 6$, may take any value. Observe that the tangent space of $\mathrm{Syl}(\sigma)$ in any matrix $M \in \mathrm{Syl}(\sigma)$ is independent of M. This is the reason why we just write $\mathrm{TSyl}(\sigma)$ without specifying the base point.

In order to extend the transversality identity (9.24) to the Sylvester space of any Fiedler matrix, we first need the following result, which is in turn an extension of [53, Eq. (5), p. 768].

Lemma 9.23. *Let E^{syl} be a matrix in $\mathrm{TSyl}(\sigma)$ with nonzero entries equal to $E_0^{\mathrm{syl}}, E_1^{\mathrm{syl}}, \ldots, E_{n-1}^{\mathrm{syl}}$, where the entry E_k^{syl}, for $k = 0, 1, \ldots, n-1$, is in the same position as the coefficient $-a_k$ in $M_\sigma(p)$ with $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$. Then, for $k = 0, 1, \ldots, n-1$,*

$$\mathrm{tr}(E^{\mathrm{syl}}\, p_{n-k-1}(M_\sigma)) = -E_k^{\mathrm{syl}}. \tag{9.25}$$

Proof. Let $\widetilde{p}(z) = z^n + \sum_{k=0}^{n-1} \widetilde{a}_k z^k$ be the characteristic polynomial of $M_\sigma + E^{\mathrm{syl}}$. We know, by Propositions 9.2 and 9.11, that $\widetilde{a}_k = a_k - \mathrm{tr}(E^{\mathrm{syl}}\, p_{n-k-1}(M_\sigma)) + O(\|E^{\mathrm{syl}}\|^2)$. But $M_\sigma + E^{\mathrm{syl}}$ is a Fiedler matrix of the polynomial $z^n + \sum_{k=0}^{n-1}(a_k + E_k^{\mathrm{syl}})z^k$, therefore we have $\widetilde{a}_k = a_k + E_k^{\mathrm{syl}}$. From these two formulas we get

$$\mathrm{tr}(E^{\mathrm{syl}}\, p_{n-k-1}(M_\sigma)) + O(\|E^{\mathrm{syl}}\|^2) = -E_k^{\mathrm{syl}}.$$

Since this last equation is true regardless of the value of $E_0^{\mathrm{syl}}, E_1^{\mathrm{syl}}, \ldots, E_{n-1}^{\mathrm{syl}}$, (9.25) follows. □

Theorem 9.24. *Let $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ be a monic polynomial, $\sigma : \{0, 1, \ldots, n-1\} \to \{1, \ldots, n\}$ be a bijection, and let M_σ be the Fiedler matrix of $p(z)$ associated to the bijection σ. Then $\mathrm{Syl}(\sigma)$ is transversal to $\mathcal{O}(M_\sigma)$ at M_σ, i.e., every matrix $E \in \mathbb{C}^{n\times n}$ can be expressed as*

$$E = E^{\tan} + E^{\mathrm{syl}}, \tag{9.26}$$

where $E^{\mathrm{syl}} \in \mathrm{TSyl}(\sigma)$ and $E^{\tan} \in T_{M_\sigma}\mathcal{O}(M_\sigma)$. Moreover, the decomposition in (9.26) is unique, that is, $\mathrm{TSyl}(\sigma) \cap T_{M_\sigma}\mathcal{O}(M_\sigma) = \{0\}$.

Proof. Let E^{syl} be a matrix in $\mathrm{TSyl}(\sigma)$ with nonzero entries $E_k^{\mathrm{syl}} := -\mathrm{tr}(E\, p_{n-k-1}(M_\sigma))$, for $k = 0, 1, \ldots, n-1$, where the entry E_k^{syl} is in the same position as $-a_k$ in M_σ. We may write the matrix E as $E^{\mathrm{syl}} + E^{\tan}$, where $E^{\tan} = E - E^{\mathrm{syl}}$. We have to check that $E^{\tan} \in T_{M_\sigma}\mathcal{O}(M_\sigma)$. Indeed, using Lemma 9.23,

$$\begin{aligned} \mathrm{tr}(E\, p_{n-k-1}(M_\sigma)) &= \mathrm{tr}(E^{\mathrm{syl}}\, p_{n-k-1}(M_\sigma)) + \mathrm{tr}(E^{\tan}\, p_{n-k-1}(M_\sigma)) \\ &= \mathrm{tr}(E\, p_{n-k-1}(M_\sigma)) + \mathrm{tr}(E^{\tan}\, p_{n-k-1}(M_\sigma)). \end{aligned}$$

From this, we deduce that $\mathrm{tr}(E^{\tan}\, p_{n-k-1}(M_\sigma)) = 0$, for $k = 0, 1, 2, \ldots, n-1$. But, from Proposition 9.21, we have that $\{p_k(M_\sigma)^*\}_{k=0}^{n-1}$ is a basis for $N_{M_\sigma}\mathcal{O}(M_\sigma)$, therefore $E^{\tan} \in T_{M_\sigma}\mathcal{O}(M_\sigma)$. □

Theorem 9.24, together with (9.25) show us that the component $E^{\tan}$ of the perturbation matrix E does not contribute to the first order term of $a_k(M_\sigma + E)$, so that only the "transversal complement" E^{syl} contributes to first order. In other words:

$$\begin{aligned} a_k(M_\sigma + E) &= a_k - \operatorname{tr}(p_{n-k-1}(M_\sigma)E)) + O(\|E\|^2) = a_k - \operatorname{tr}(p_{n-k-1}(M_\sigma)E^{\text{syl}}) + O(\|E\|^2) \\ &= a_k(M_\sigma + E^{\text{syl}}) + O(\|E\|^2). \end{aligned}$$

Also, from the considerations above, if E_k^{syl} denotes, as in Lemma 9.23, the entry of E^{syl} which is located in the same position as the coefficient $-a_k$ in M_σ, then we have, up to first order in E,

$$E_k^{\text{syl}} = a_k(M_\sigma + E) - a_k = -\sum_{i,j=1}^{n} p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})E_{ij}, \tag{9.27}$$

as in (9.3), with $p_{ij}^{(\sigma,k)}(a_0, a_1, \ldots, a_{n-1})$ given by Theorem 9.3. Recall that the remaining entries of E^{syl} are zero. Hence, from (9.26) and (9.27) we may get explicit expressions for the entries of $E^{\tan} = E - E^{\text{syl}}$ in terms of the entries of E and the coefficients $a_0, a_1, \ldots, a_{n-1}$.

We want to emphasize that in the approach followed by [53], the fact that E^{syl} is transversal to the tangent space of $\mathcal{O}(C_1)$ at C_1 is key to get the first order expression for $a_k(C_1 + E)$. More precisely: using this transversality (namely, equation (9.26) with E^{syl} being the Sylvester space for C_1), together with the identity $\operatorname{tr}(p_{n-k}(C_1)E^{\tan}) = 0$, and the explicit expression $-E_{k-1}^{\text{syl}} = \operatorname{tr}(p_{n-k}(C_1)E)$, both them valid for $k = 1, \ldots, n$, they get an explicit expression for $\operatorname{tr}(p_{n-k}(C_1)E)$, which is the first order term of $a_{k-1}(C_1 + E)$. This can be done because the matrices $p_{n-k}(C_1)$, for $k = 1, \ldots, n$, have a simple structure that allows to compute $\operatorname{tr}(p_{n-k}(C_1)E^{\text{syl}})$ easily and explicitly, for all $k = 1, \ldots, n$. Unfortunately, for arbitrary Fiedler matrices, to get explicit expressions of $\operatorname{tr}(p_{n-k}(M_\sigma)E)$ by hand is quite involved. Hence, we have obtained the first-order term of $a_k(M_\sigma + E)$ directly from $\operatorname{adj}(zI - M_\sigma)$. This approach is completely independent of the transversality of E^{syl} and the tangent space, though, as we have seen in Theorem 9.24, this fact is still true for arbitrary Fiedler matrices.

Chapter 10

Conclusions, publications, and open problems

In this chapter we summarize the main original contributions of this dissertation, we discuss some related work, and list the papers published or submitted containing most of the results presented in this thesis. In addition, we propose a set of related open problems for future research.

10.1 Conclusions and original contributions

Chapter 3: We have shown how to construct the inverses of Fiedler companion matrices and we have studied some of their properties. We have also obtained explicit expressions of the norms of Fiedler matrices and their inverses in the case of the 1-, ∞-, and Frobenius matrix norms.

Chapter 4: We have performed a study of singular values of Fiedler companion matrices of a monic polynomial $p(z)$. We have seen that the singular values of Frobenius companion matrices have very simple properties that are not shared by any other Fiedler matrix. Nonetheless, the singular values of Fiedler matrices still retain some interesting properties that we have carefully studied, that is, we have determined how many singular values of a Fiedler matrix are equal to one, and, for those that are not, we have showed that they can be obtained from the square roots of the eigenvalues of certain matrices that have a size much smaller than the degree of $p(z)$ and that are easily constructible from the coefficients of $p(z)$. This study is based on the developments that we have presented on a new class of matrices termed as "staircase matrices", which have a very special zero pattern.

Chapter 5: We have obtained two different explicit expressions for the adjugate matrix of $zI - M_\sigma$, where M_σ is a Fiedler companion matrix. These expressions have been later used in Chapters 8 and 9.

Chapter 6: Explicit expressions and a complete analysis of the bounds on the absolute values of the roots of a monic scalar polynomial that are obtained by using the 1-, ∞-, and Frobenius norms of Fiedler companion matrices and their inverses have been presented in this chapter. Particular attention has been paid to determine which are the sharpest bounds among those coming from Fiedler matrices and their inverses, and we have found that in many interesting situations the bounds coming from the inverse of the Fiedler matrix F defined in (2.8):

$$\min\left\{\frac{|a_0|}{1+|a_1|}, \frac{1}{1+|a_2|}, \ldots, \frac{1}{1+|a_{n-1}|}\right\} \leq |\lambda| \leq \max\left\{1+\frac{|a_1|}{|a_0|}, 1+\frac{|a_2|}{|a_0|}, \ldots, 1+\frac{|a_{n-2}|}{|a_0|}, |a_0|+|a_{n-1}|\right\},$$

where λ is a root of $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$, are the sharpest ones and that they improve significatively, for certain polynomials, the classical bounds obtained from the Frobenius companion matrices.

Chapter 7: We have performed a very detailed study of the condition numbers for inversion of Fiedler companion matrices of monic polynomials $p(z)$ in the Frobenius norm. This study is based on the new properties for the inverses of Fiedler companion matrices obtained in Chapter 3. We have established that, from the point of view of condition numbers for inversion, the classical Frobenius companion matrices should not be used if $|p(0)| < 1$, since they have the largest condition number among all the Fiedler matrices of $p(z)$ and one should use, instead, any Fiedler matrix having a number of initial consecutions or inversions equal to 1. On the contrary, if $|p(0)| > 1$, then the Frobenius companion matrices are the ones to be used, since they have the smallest condition number among all the Fiedler matrices of $p(z)$. In the border case $|p(0)| = 1$ all Fiedler matrices of p(z) have the same condition number. However we have also established that, given a monic polynomial $p(z)$, if there are two distinct Fiedler matrices with very different condition numbers, then both matrices are very ill-conditioned. Therefore, different Fiedler matrices may have very different condition numbers but only in cases where these matrices are nearly singular. Loosely speaking, this means that there is no any polynomial $p(z)$ for which one Fiedler matrix has a small condition number while others have very large condition numbers.

Chapter 8: We have carried out a detailed study of the eigenvalue condition numbers of Fiedler companion matrices of a monic polynomial $p(z)$. Ideally, in the polynomial root-finding problem using Fiedler companion matrices, one would like the eigenvalues of the Fiedler matrix to be as well conditioned as the roots of the original polynomial:

$$\frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} = \Theta(1),$$

where $\kappa(\lambda, M_\sigma)$ and $\kappa(\lambda, p)$ denote, respectively, the condition number of λ as an eigenvalue of M_σ and the condition number of λ as a root of $p(z)$. However, we have seen that

$$\frac{1}{\sqrt{2}} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq n\rho(p)\|p\|_2$$

if $M_\sigma = C_1, C_2$, and

$$\frac{1}{n} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, p)} \leq n^{5/2}\rho(p)\|p\|_2^2,$$

if $M_\sigma \neq C_1, C_2$, where

$$\|p\|_2 = \sqrt{1 + \sum_{k=0}^{n-1} |a_k|^2} \quad \text{and} \quad \rho(p) = \sqrt{1 + \frac{1}{\max_{0\leq k\leq n-1} |a_k|^2}}.$$

These bounds have led us to conclude that, from the point of view of eigenvalue condition numbers, any Fiedler matrix can be used for solving the root-finding problem for $p(z)$ when the absolute value of the coefficients of $p(z)$ are moderate and not close to zero. On the contrary, when $\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}$ is large or close to zero, the eigenvalues of any Fiedler companion matrix may be potentially more ill conditioned than the roots of $p(z)$.

We have also studied the ratio between the eigenvalue condition numbers of Fiedler matrices other that the Frobenius ones and the eigenvalue condition number of Frobenius companion matrices, that is, the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$, where C denotes the first or the second Frobenius companion matrices. We have proved that

$$(n^2\|p\|_2)^{-1} \leq \frac{\kappa(\lambda, M_\sigma)}{\kappa(\lambda, C)} \leq n^{5/2}\|p\|_2,$$

which allows us to conclude that, from the point of view of eigenvalue condition numbers, when $\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}$ is moderate, any Fiedler matrix can be used for solving the root-finding problem for $p(z)$ with the same reliability as Frobenius companion matrices. On the other hand, when $\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}$ is large, we have shown that the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ may be arbitrarily large or arbitrarily small. In addition to this, we have shown that if this ratio is very large, then λ is very ill conditioned as an eigenvalue of M_σ and as an eigenvalue of the Frobenius companion matrices compared to $\kappa(\lambda, p)$. On the other hand, we have also shown that the opposite is not true. There exist polynomials for which the ratio $\kappa(\lambda, M_\sigma)/\kappa(\lambda, C)$ is arbitrarily small, but this only implies that λ is very ill-conditioned as an eigenvalue of the Frobenius companion matrices compared with $\kappa(\lambda, p)$. From the point of view of eigenvalue condition numbers, this allows to conclude that there are polynomials for which one should avoid computing their roots as the eigenvalues of Frobenius companion matrices and to use, instead, another Fiedler matrix. Although how to identify these polynomials and how to know which Fiedler matrix one might use instead of the Frobenius ones is an interesting open problem in this area.

Regarding pseudospectra of Fiedler matrices, we have shown how to estimate accurately them in a $m \times m$ grid using only $O(nm^2)$ flops compare with the $O(n^3 + n^2m^2)$ flops needed for general matrices (see Section 1.2.2.2). Then, we have established various mathematical relationships between the pseudozero sets of a monic polynomial $p(z)$ and the pseudospectra of the associated Fiedler matrices which have led us to reach the same conclusions that we have stated in the two previous paragraphs.

Finally, we have also studied numerically the effect of balancing Fiedler companion matrices on the eigenvalue condition numbers and pseudospectra. We have provided numerical experiments that show that the eigenvalues of Fiedler matrices that have been previously balanced and the root of monic polynomials are essentially equally conditioned, generically.

Chapter 9: We have analyzed the backward stability of the polynomial root-finding problem when considered as a standard eigenvalue problem by means of Fiedler companion matrices. For this purpose, we have described the first-order change of the characteristic polynomial of any Fiedler matrix under small perturbations of the matrix. This description has led us to conclude that polynomial root-finding algorithms based on backward stable eigenvalue algorithms using Fiedler companion matrices, are backward stable from the point of view of the polynomials only if $\|p\|_\infty$ is moderate. More precisely, given a monic polynomial $p(z)$, if $\widetilde{p}(z)$ denotes the monic polynomial whose roots are the computed eigenvalues of a Fiedler companion matrix of $p(z)$, obtained with a backward stable eigenvalue algorithm, then it is not possible to guarantee, in general, that

$$\frac{\|\widetilde{p} - p\|_\infty}{\|p\|_\infty} = O(u),$$

where u is the machine epsilon of the computer. Namely, the computed roots of $p(z)$ are not necessarily the roots of a nearby polynomial. We have seen, however, that

$$\frac{\|\widetilde{p} - p\|_\infty}{\|p\|_\infty} = O(u)\|p\|_\infty^2,$$

for any Fiedler companion matrix other than the first and second Frobenius companion matrices, and that

$$\frac{\|\widetilde{p} - p\|_\infty}{\|p\|_\infty} = O(u)\|p\|_\infty,$$

for the first and second Frobenius companion matrices (which are particular cases of Fiedler matrices). These bounds allow us to conclude that, from the point of view of the backward

errors, any Fiedler matrix can be used for solving the root-finding problem with the same reliability as Frobenius companion matrices when the absolute value of the coefficients of $p(z)$ are moderate. In this case, as we said, the root-finding problem solved by applying a backward stable eigenvalue algorithm on any Fiedler companion matrix is a backward stable method. On the other hand, when $p(z)$ has large coefficients any Fiedler matrix other than Frobenius companion matrices may produce much larger backward errors than the ones produced when using Frobenius matrices, but, in this situation, none of the Fiedler matrices (including the Frobenius ones) leads to a backward stable algorithm for the rootfinding problem. Extensive numerical experiments have been included to confirm these theoretical results.

We have also studied the effect of balancing Fiedler companion matrices on the backward errors of the root-finding problem for $p(z)$ using a balanced Fiedler matrix M_σ. The numerical experiments that we have carried out in Chapter 9 indicate that balancing very often improves the backward errors for general polynomials. In fact, balancing Fiedler matrices is usually enough to guarantee that the process of computing the roots of a polynomial as the eigenvalues of a Fiedler matrix is backward stable in the polynomial sense, even if the polynomial has large coefficients. However, we have also proved that there are infinitely many polynomials for which balancing the Fiedler companion matrices does not guarantee backward stability for the root-finding polynomial problem.

In summary, we have seen that Fiedler companion matrices present a very rich structure and very interesting algebraic and numerical properties. Most of these properties are similar to those of the Frobenius companion matrices. However, we have also shown that there are relevant differences between some properties of Frobenius companion matrices and those of the rest of Fiedler companion matrices (in particular, properties regarding norms of their inverses, condition numbers for inversion, eigenvalue condition numbers, and backward errors), although we have seen that to study these properties of Fiedler companion matrices is much more complicated than to study them in the case of the Frobenius companion matrices. In addition, the results in Chapter 7, 8 and 9 allow us to conclude that any Fiedler companion matrix can be used in numerical applications with the same reliability as Frobenius companion matrices in the same situations where Frobenius companion matrices can be used in a reliable way. Therefore, in those cases, one can take advantage of the particular structure of some Fiedler matrices to make their use more efficient than the use of classical Frobenius companion matrices. For instance, one could take advantage of the pentadiagonal structure of the Fiedler matrices in (2.7) to devise structured versions of the LR algorithm to get its eigenvalues in $O(n^2)$ flops.

10.2 Publications

The original results contained in this dissertation have been published in several international research Journals, all of them indexed in the *Journal Citation Reports of ISI Web of Knowledge*.

The results in Chapters 3, 4 and 7 are contained in:

F. De Terán, F. M. Dopico, and J. Pérez. Condition numbers for inversion of Fiedler companion matrices. *Linear Algebra and its Applications*, 439, pp. 944–981, 2013.

The results in Chapters 3 and 6 are contained in:

F. De Terán, F. M. Dopico, and J. Pérez. New bounds for roots of polynomials based on Fiedler companion matrices. *Linear Algebra and its Applications*, 451, pp. 197–230, 2014.

The results in Chapters 5 and 9 are contained in:

> F. De Terán, F. M. Dopico, and J. Pérez. Backward stability of polynomial root-finding using Fiedler companion matrices. *Accepted in IMA Journal of Numerical Analysis.*

The results in Chapter 8 will be submitted to publication shortly.

10.3 Open problems

Finally, we discuss some related work as well as we propose a set of open problems for future research related to the problems solved in this dissertation.

P1. Singular values of Fiedler companion matrices: At present, there are no explicit expressions for the singular values of those Fiedler matrices that are different from the Frobenius ones. We have determined how many of their singular values are exactly equal to one and, for those that are not, we have showed that they can be obtained from the square roots of the eigenvalues of certain matrices that have a size much smaller than n, although it is not known how to get the explicit expressions of the singular values from those small matrices. These expressions could be used to perform a study of the condition number for inversion of Fiedler matrices similar to the one presented in this dissertation but using the spectral norm instead of the Frobenius norm. More important, they could be used also to get new and maybe tighter upper and lower bounds for the absolute values of the roots of monic polynomials. Therefore, to get explicit expressions for these singular values or at least good simple approximations for them is an interesting open problem in this area.

P2. New bounds for roots of scalar polynomials and new bounds for eigenvalues of matrix polynomials: The work presented in Chapter 6 is just a first step in the use of Fiedler matrices for bounding roots of polynomials. Next steps will include:

(a) the generalization of the results presented in Chapter 6 from scalar to matrix polynomials, since Fiedler companion matrices have been extended, and thoroughly studied, to the context of matrix polynomials [5, 45, 47];

(b) the investigation of concrete diagonal scalings of Fiedler matrices and/or their inverses that can produce sharper bounds for some classes of scalar polynomials; and

(c) the use of Fiedler matrices for getting other types of inclusion regions for the roots of scalar polynomials, as it was done in [119] for the classical Frobenius companion matrices.

P3. Backward stability for the root-finding problem of non-monic polynomials using Fiedler companion pencils One way to circumvent the inaccuracies due to the occurrence of large polynomial coefficients is to shift from companion matrices to companion pencils where normalization can be applied (see [93]). Though exactly the same techniques used in [93] for the Frobenius companion pencils can not be directly applied to other Fiedler companion pencils, some further analysis in this direction is still to be done, and will be the subject of future work. A possible approach may be to use similar techniques to the ones in [160], where the authors prove that solving a matrix polynomial eigenvalue problem by applying the QZ algorithm to the Frobenius companion pencil is backward stable from the point of view of the polynomials, provided that the original matrix polynomial has been previously scaled so that all coefficients have norm

less than or equal to 1, since these techniques has been used in [126] to prove the backward stability of algorithms that compute the roots of polynomials via the eigenvalues of comrade pencils.

P4. Backward error of polynomial eigenproblems solved by Fiedler companion pencils: Backward errors of single roots has been considered in [147] for the more general case of matrix Polynomial Eigenvalue Problems. In particular, the backward error of a single computed eigenvalue $\widetilde{\lambda}$ of a matrix polynomial $P(\lambda)$ considered in [147] is:

$$\eta(\widetilde{\lambda}) = \min\left\{\epsilon : \ (P + \Delta P)(\widetilde{\lambda}) = 0, \quad \|\Delta A_i\|_2 \leq \epsilon \|A_i\|_2, \ i = 0, 1, \ldots, n\right\},$$

where $P(\lambda) = \sum_{k=0}^{n} A_k \lambda^k$, and $\Delta P(\lambda) = \sum_{k=0}^{n} (\Delta A_k)\lambda^k$ are not necessarily monic. It is shown in [82] that, for matrix polynomials all whose coefficients have 2-norm not far from 1, computing the eigenvalues of its Frobenius companion pencil with a backward stable eigenvalue algorithm gives a coefficientwise backward stable method for the Polynomial Eigenvalue Problem, that is, each computed eigenvalue is the exact eigenvalue of a nearby matrix polynomial, but this matrix polynomial need not necessarily to be the same for each eigenvalue. The problem of proving that solving matrix Polynomial Eigenvalue Problems by applying a backward stable eigenvalue algorithm to the Frobenius companion pencil is backward stable in the sense that the *whole ensemble* of computed eigenvalues is the whole ensemble of exact eigenvalues of a nearby matrix polynomial, was solved in [160]. In that paper, the authors prove that if the norms of the matrix coefficients are bounded by one, then the whole ensemble of computed eigenvalues is the whole ensemble of exact eigenvalues of a nearby matrix polynomial in the normwise sense $\|\Delta A_i\|_2 \leq u\|[A_0, A_1, \ldots, A_n]\|$, but not nearby in a coefficientwise sense $\|\Delta A_i\| \leq u\|A_i\|_2$, where u is the unit roundoff. This problem is still open when Fiedler companion pencils other than the Frobenius ones are used instead of the Frobenius companion pencils, and will be the subject of future work.

P5. Backward error of polynomial eigenproblems solved by generalized Fiedler pencils and Fiedler pencils with repetitions: As we commented in Section 1.5, finding linearizations that retain whatever structures that a matrix polynomial $P(\lambda)$ might possess has motivated in the last few years an intense activity on the development of new classes of linearizations. Based on Fiedler pencils, two classes of linearizations that contain structure preserving linearizations has been introduced: generalized Fiedler pencils [5, 6, 30], and Fiedler pencils with repetitions [29, 31, 32, 33, 164]. To extend the backward error analysis to polynomial eigenproblems solved by generalized Fiedler pencils and Fiedler pencils with repetition will be the subject of future work.

Bibliography

[1] O. Alberth. Iteration methods for finding all zeros of a polynomial simultaneously. *Math. Comp.*, 27, pp. 339–344, 1973.

[2] J. Ackermann. Der entwurf linearer regelungssysteme im zustandsraum, *Regelungtechnik und Prozessdatenverarbeitung*, 7, pp. 297–300, 1972

[3] A. Amiraslani, D. A. Aruliah, R. M. Corless. Block LU factors of generalized companion matrix pencils. *Theor. Comp. Sci.*, 381, pp. 134–147, 2007.

[4] A. Amiraslani, R. M. Corless, P. Lancaster. Linearization of matrix polynomials expressed in polynomial bases. *IMA J. Numer. Anal.*, 29, pp. 141–157, 2009.

[5] E. N. Antoniou, S. Vologiannidis. A new family of companion forms of polynomial matrices. *Electron. J. Linear Algebra*, 11, pp. 78–87, 2004.

[6] E. N. Antoniou, S. Vologiannidis. Linearizations of polynomial matrices with symmetries and their applications. *Electron. J. Linear Algebra*, 15, pp. 107–114, 2006.

[7] V. I. Arnold. On matrices depending on parameters. *Russian Math. Surveys*, 26, pp. 29–43, 1971.

[8] J. L. Aurentz, T. Mach, R. Vandebril, D. S. Watkins. Fast and backward stable computation of roots of polynomials. TW654, Department of Computer Science, KU Leuven, Leuven, Belgium, 2014.

[9] J. L. Aurentz, R. Vandebril, D. S. Watkins. Fast computation of the zeros of a polynomial via factorization of the companion matrix. *SIAM J. Sci. Comput.*, 35, pp. 255–269, 2013

[10] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: a Practical Guide*. SIAM, Philadelphia, 2000.

[11] W. W. Barret, P. J. Feinsilver. Gaussian families and a theorem of patterned matrices. *J. Appl. Probab.*, 15, pp. 514-522, 1978.

[12] C. Bekas, E. Gallopoulos. Cobra: parallel path following for computing the matrix pseudospectrum. *Parallel Computing*, 27, pp. 1879–1896, 2001.

[13] C. Bekas, E. Gallopoulos. Parallel computation of pseudospectra by fast descent. *Parallel Computing*, 28, pp. 223–242, 2002.

[14] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas.* Princeton University Press, Princeton, 2009.

[15] R. Bhatia, T. Jain. Higher order derivatives and perturbation bounds for the determinant. *Linear Algebra Appl.*, 431, pp. 2102–2108, 2009.

[16] D. A. Bini, L. Gemignani, F. Tisseur. The Ehrlich-Aberth method for the non-symmetric tridiagonal eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 27(1), pp. 153–175, 2005.

[17] D. A. Bini. Numerical computation of polynomial zeros by means of Aberth's method. *Numer. Algorithms*, 13, pp. 179–200, 1996.

[18] D. A. Bini, G. Fiorentino. Design, analysis, and implementation of a multiprecision polynomial rootfinder. *Numer. Algorithms*, 23(2-3), pp. 127–173, 2000.

[19] D. A. Bini, L. Gemignani, V. Y. Pan. Improved initialization of the accelerated and robust QR-like polynomial root-finding. *Electron. Trans. Numer. Anal.*, 17, pp. 195–205, 2004.

[20] D. A. Bini, L. Gemignani, V. Y. Pan. Fast and stable QR eigenvalue algorithms for generalized companion matrices and secular equations. *Numer. Math.*, 100, pp. 373–408, 2006

[21] D. A. Bini, L. Gemignani, V. Y. Pan. QR-like algorithm for generalized semiseparable matrices. Tech. Report 1470, Department of Mathematics, University of Pisa, Italy, 2004.

[22] D. A. Bini, P. Boito, Y. Eidelman, L. Gemignani, I. Gohberg. A fast implicit QR eigenvalue algorithm for companion matrices. *Linear Algebra Appl.*, 432, pp. 2006–2031, 2010.

[23] D. A. Bini, V. Noferini, M. Sharify. Locating the eigenvalues of matrix polynomials. *SIAM. J. Matrix Anal. Appl.*, 34(4), pp. 1708-1727, 2013.

[24] R. R. Bitmead, H. Weiss. On the solution of the discrete-time Lyapunov matrix equation in controllable canonical form. *IEEE Trans. Automat. Control*, 24, pp. 481–482, 1979.

[25] P. Boito, Y. Eidelman, L. Geminani. Implicit QR for rank-structured matrix pencils. *B*IT Numer. Math., 54(1), pp. 85–111, 2014.

[26] C. A. Boyer. *A History of Mathematics.* Wiley, 1969.

[27] W. L. Brogan, Application of a determinant identity to pole-placement and observer problems. *IEEE Trans. Automat. Control*, 19, pp. 612–614, 1974.

[28] L. Brugnano, D. Trigiante. (1995) Polynomial roots: the ultimate answer? *Linear Algebra Appl.*, 225, 207–219, 1995.

[29] M. I. Bueno, F. De Terán. Eigenvectors and minimal bases for some families of Fiedler like linearizations. *Lin. Multilin. Algebra,* 62(1), pp. 39–62, 2014.

[30] M. I. Bueno, F. De Terán, F. M. Dopico. Recovery of eigenvectors and minimal bases of matrix polynomials from generalized Fiedler linearizations. *SIAM J. Matrix Anal. Appl.*, 32, pp. 463–483, 2011.

[31] M. I. Bueno, S. Furtado. T-Palindromic linearizations of a matrix polynomial of odd degree obtained from Fiedler pencils. *Electron. J. Linear Algebra*, 23, pp. 562–577, 2012.

[32] M. I. Bueno, K. Curlett, S. Furtado. Structured strong linearizations from Fiedler pencils with repetition I. *To appear in Linear Algebra appl.*

[33] M. I. Bueno, S. Furtado. Structured strong linearizations from Fiedler pencils with repetition II. *To appear in Linear Algebra appl.*

[34] D. Calvetti, S.- M. Kim, L. Reichel. The restarted QR-algorithm for eigenvalue computation of structured matrices. *J. Comput. Appl. Math.*, 149, pp. 415–422, 2002.

[35] A. L. Cauchy. *Exercises de Mathématiques.* 1829, Euvres, 9(2), p. 122.

[36] S. Chandrasekaran, M. Gu, J. Xia, J. Zhu. A fast QR algorithm for companion matrices. *Oper. Theory Adv. Appl.*, 179, pp. 111–143, 2007.

[37] C. L. Cox, W. F. Moss. Backward error analysis for a pole assignment algorithm. *SIAM J. Matrix Anal. Appl.*, 10, pp. 446–456, 1989.

[38] C. L. Cox, W. F. Moss, Backward error analysis for a pole assignment algorithm II: the complex case. *SIAM J. Matrix Anal. Appl.*, 13, pp. 1159–1171, 1992.

[39] J. Kautsky, N. K. Nichols, P. Van Dooren. Robust pole assignment in linear state feedback. *Internat. J. Control*, 41, pp. 1129–1155, 1985.

[40] S. Delvaux, M. Van Barel. A QR-based solver for rank structured matrices. *SIAM J. Matrix Anal. Appl.*, 30, pp. 164–490, 2008.

[41] J. W. Demmel. *Applied Numerical Linear Algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[42] J. W. Demmel, B. Kågström. The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: robust software with error bounds and applications. Part I: theory and algorithms. *ACM Trans. Math. Software*, 19, pp. 160–174, 1993.

[43] J. W. Demmel, B. Kågström. The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: robust software with error bounds and applications. Part II: software and applications. *ACM Trans. Math. Software*, 19, pp. 175–201, 1993.

[44] F. De Terán, F. M. Dopico, D. S. Mackey. Linearizations of singular matrix polynomials and the recovery of minimal indices. *Electron. J. Linear Algebra*, 18, pp. 371–402, 2009.

[45] F. De Terán, F. M. Dopico, D. S. Mackey. Fiedler companion linearizations and the recovery of minimal indices. *SIAM J. Matrix Anal. Appl.*, 31, pp. 2181–2204, 2010.

[46] F. De Terán, F. M. Dopico, D. S. Mackey. Palindromic companion forms for matrix polynomials of odd degree. *J. Comput. Appl. Math.*, 236, pp. 1464–1480, 2011.

[47] F. De Terán, F. M. Dopico, D. S. Mackey. Fiedler companion linearizations for rectangular matrix polynomials. *Linear Algebra Appl.*, 437, pp. 957–991, 2012.

[48] F. De Terán, F. M. Dopico, D. S. Mackey. Spectral equivalence of matrix polynomials and the index sum theorem. *Linear Algebra Appl.*, 459, pp. 264–333, 2014.

[49] F. De Terán, F. M. Dopico, J. Pérez. Condition numbers for inversion of Fiedler companion matrices. *Linear Algebra Appl.*, 439, pp. 944–981, 2013.

[50] F. De Terán, F. M. Dopico, J. Pérez. New bounds for roots of polynomials based on Fiedler companion matrices. *Linear Algebra Appl.*, 451, pp. 197–230, 2014.

[51] F. De Terán, F. M. Dopico, and J. Pérez. Backward stability of polynomial root-finding using Fiedler companion matrices. *IMA J. Numer. Anal.*, in press. doi: 10.1093/imanum/dru057.

[52] F. De Terán, F. Tisseur. Backward error and conditioning of Fiedler linearizations. In progress.

[53] A. Edelman, H. Murakami. Polynomial roots from companion matrix eigenvalues. *Math. Comp.*, 210, pp. 763–776, 1995.

[54] A. Edelman, E. Elmroth, B. Kågström. A geometric approach to perturbation theory of matrices and matrix pencils. Part I: versal deformations. *SIAM J. Matrix Anal. Appl.*, 18, pp. 653–692, 1997.

[55] A. Edelman, E. Elmroth, B. Kågström. A geometric approach to perturbation theory of matrices and matrix pencils. Part II: a stratifications-enhanced staircase algorithm. *SIAM J. Matrix Anal. Appl.*, 20, pp. 667–699, 1999.

[56] L. W. Ehrlich. A modified Newton method for polynomials. *Comm. ACM*, 10, pp. 107–108, 1967.

[57] Y. Eidelman, I. Gohberg, V. Olshevsky. The QR iteration method for hermitian quasiseparable matrices of arbitrary order. *Linear Algebra Appl.*, 404, pp. 305–324, 2005.

[58] F. Fiedler. Structured ranks of matrices. *Linear Algebra Appl.*, 197, pp. 119–127, 1993.

[59] M. Fiedler. A note on companion matrices. *Linear Algebra Appl.,* 372, pp. 325–331, 2003.

[60] M. Fiedler Complementary basic matrices. *Linear Algebra Appl.*, 384, pp. 199–206, 2004.

[61] R. A. Frazer, W. J. Duncan, A. R. Collar. *Elementary matrices.* Cambridge University Press, Cambridge, 1965.

[62] G. D. Forney. Minimal bases of rational vector spaces, with applications to multivariable linear systems. *SIAM J. Control*, 13(3), pp. 493–520, 1975.

[63] V. Frayssé, M. Gueury, F. Nicoud, V. Toumazou. Spectral Portraits for Matrix Pencils. Technical Report TR/PA/96/19, CERFACS, Toulouse, France, 1996.

[64] G. Frobenius. Theorie der linearen formen mit ganzen coefficienten. *J. Reine Angew. Math.*, 86, pp. 146–208, 1879.

[65] M. Fujiwara. Uber die obere schranke des absoluten betrages der wurzeln einer algebraischen gleichung. *Tohoku Math. J.*, 10, pp. 167–171, 1916.

[66] F. R. Gantmacher. *Theory of Matrices.* Chelsea, New York, 1959.

[67] F. R. Gantmacher, M. G. Krein. *Oscillations matrices and kernels and small vibrations of mechanical systems.* AMS Chelsea Publishing, Providence, 2002.

[68] E. Gallestey. Computing the spectral value sets using the subharmonicity of the norm of rational matrices. *BIT*, 38, pp. 22–33, 1998.

[69] L. Gemignani. Structured matrix methods for polynomial root-finding. *Proceedings ISSAC*, pp. 175–180, 2007.

[70] I. Gohberg, T. Kailath, I. Koltracht. Linear complexity algorithms for semiseparable matrices. *Integr. Eq. Oper. Theory*, 8(6), pp. 780–804, 1985.

[71] I. Gohberg, P. Lancaster, L. Rodman. *Matrix Polynomials.* Academic Press, New York, 1982.

[72] I. Gohberg, M. A. Kaashoek, P. Lancaster. General theory of regular matrix polynomials and band Toeplitz operators. *Integr. Eq. Oper. Theory*, 11, pp. 776–882, 1988.

[73] G. Golub, C. Van Loan. *Matrix Computations.* Third Ed., Johns Hopkins University Press, Baltimore, 1996.

[74] H. Grauert, K. Fritzsche. *Several Complex Variables.* Springer, Berlin, 1976.

[75] C. He, A. J. Laub, V. Mehrmann, Placing plenty of poles is pretty preposterous. *Syst. Control Lett.*, 26, pp. 275–281, 1995.

[76] N. J. Higham, F. Tisseur. Bounds for eigenvalues of matrix polynomials. *Linear Algebra Appl.*, 358, pp. 5–22, 2003.

[77] N. J. Higham, F. Tisseur. More on pseudospectra for polynomial eigenvalue problems and applications in control theory. *Linear Algebra Appl.*, 351-352, pp. 435-453, 2002.

[78] D. J. Higham, L. N. Trefethen. Stiffness of ODEs. *BIT*, 33, pp. 285-303, 1993.

[79] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2002.

[80] N. J. Higham, D. S. Mackey, F. Tisseur. The conditioning of linearizations of matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 28, pp. 1005–1024, 2006.

[81] N. J. Higham, D. S. Mackey, N. Mackey, F. Tisseur. Symmetric linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 29, pp. 143–159, 2006

[82] N. J. Higham, R-C. Li, F. Tisseur. Backward error of polynomial eigenproblems solved by linearization. *SIAM J. Matrix Anal. Appl.*, 29, pp. 1218–1241, 2007.

[83] N. J. Higham, D. S. Mackey, F. Tisseur, S. D. Garvey. Scaling, sensitivity and stability in the numerical solution of quadratic eigenvalue problems. *Int. J. Numer. Meth. Eng.*, 73, pp. 344–360, 2008.

[84] D. Hinrichsen, B. Kelb. Spectral value sets: a graphical tool for robustness analysis. *Systems and Control Lett.* 21, pp. 127–136, 1993.

[85] D. Hinrichsen, A. J. Pritchard. On spectral variations under bounded real matrix perturbations. *Numer. Math.*, 60, pp. 509–524, 1992.

[86] D. Hinrichsen, A. J. Pritchard. Stability of uncertain systems. *Systems and Networks: Mathematical Theory and Applications*, Vol. 1 (Regensburg, 1993), Math. Res. 77, Akademie-Verlag, Berlin, 1994.

[87] R. A. Horn, C. R. Johnson. *Matrix Analysis.* Second Ed. Cambridge University Press, Cambridge, 2013.

[88] R. A. Horn, C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, New York, 1991.

[89] I. Ipsen, R. Rehman. Perturbation bounds for determinants and characteristic polynomials. *SIAM J. Matrix Anal. Appl.*, 30, pp. 762–776, 2008.

[90] M. A. Jenkins, J. F. Traub. A three-stage algorithm for real polynomials using quadratic iteration. *SIAM J. Numer. Anal. Appl.*, 7, pp. 545-566, 1970.

[91] M. A. Jenkins, J. F. Traub. Principles for testing polynomial zerofinding programs. Tech. Report, Department of Computer Science, Carnegie-Mellon University, 1974.

[92] G. F. Jó nsson, L. N. Trefethen. A numerical analyst looks at the "cutoff phenomenon" in card shuffling and other Markov chains. In D. F. Griffiths, D. J. Higham and G. A. Watson, eds., *Numerical Analysis* 1997 (Dundee, 1997), 150-178. Longman Sci. Tech., Harlow, Essex, UK, 1998.

[93] G. F. Jonsson, S. Vavasis. Solving polynomials with small leading coefficients. *SIAM J. Numer. Anal. Appl.*, 26, pp. 404–414, 2004.

[94] T. Kailath. *Linear Systems*. Prentice Hall, Engewood Cliffs, 1980.

[95] C. Kenney, A. J. Laub. Controllability and stability radii for companion form systems. *Math. Control Signals Systems*, 1, pp. 239–256, 1988.

[96] O. Kerner. Ein gesamtschrittverfahren zur berechnung der nullstellen von polynomen. *Numer. Math.*, 8, pp. 290–294, 1966.

[97] T. Kojima. On the limits of the roots of an algebraic equation. *Tohoku Math. J.*, 11, pp. 119–127, 1917.

[98] I. Koutis, E. Gallopoulos. Exclusion regions and fast estimation of pseudospectra. Tech. report, Department of Computer Engineering and Informatics, HPCLAB, University of Patras, Patras, Greece, 2000.

[99] A. J. Laub. *Matrix Analysis for Scientists & Engineers*. SIAM, Philadelphia, 2005.

[100] P. W. Lawrence, R. M. Corless. Stability of rootfinding for barycentric Lagrange interpolants. *Numer. Algorithms*, 65, pp. 447–464, 2014.

[101] D. Lemmonier, P. Van Dooren. Optimal scaling of companion pencils for the QZ-algorithm. *Proceedings SIAM Appl. Lin. Alg. Conference*, paper CP7–4, 2003.

[102] D. Lemmonier, P. Van Dooren. Optimal scaling of block companion pencils. *Proceedings of the International Symposium on Mathematical Theory of Networks and Systems*, Leuven, Belgium, 2004.

[103] S. H. Lui. Computation of pseudospectra by continuation. *SIAM J. Sci. Comput.* 18, pp. 565-573, 1997.

[104] D. S. Mackey. The continuing influence of Fiedler's work on companion matrices. *Linear Algebra Appl.*, 439(4), pp. 810–817, 2013.

[105] D. S. Mackey, N. Mackey, C. Mehl, V. Mehrmann. Vector spaces of linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 28, pp. 971–1004, 2006.

[106] D. S. Mackey, N. Mackey, C. Mehl, V. Mehrmann. Structured polynomial eigenvalue problems: good vibrations from good linearizations. *SIAM J. Matrix Anal. Appl.*, 28, pp. 1029–1051, 2006.

[107] K. Madsen. A root-finding algorithm based on Newton's method. *BIT*, 13, pp. 71–75, 1973.

[108] M. Marden. *Geometry of Polynomials*. Mathematical Surveys, vol. 3, Second Ed., AMS, Providence, USA, 1966.

[109] D. Manocha, J. Demmel. Algorithms for intersecting parametric and algebraic curves (ii): simple intersections. *ACM Trans. Graphic*, 13, pp. 73–100, 1994.

[110] D. Manocha, J. Demmel. Algorithms for intersecting parametric and algebraic curves (i): higher order intersections. *Computer vision, graphics and image processing: graphical models and Image Precessing*, 57, pp. 80–100, 1995.

[111] P. G. Martinsson, V. Rokhlin. A fast direct solver for boundary integral equations in two dimensions. *J. Comput. Phys.*, 205, pp. 1–23, 2005.

[112] P. G. Martinsson, V. Rokhlin, M. Tygert. A fast algorithm for the inversion of general Toeplitz matrices. *Comput. Math. Appl.*, 50, pp. 741–752, 2005.

[113] The MathWorks, Inc. *MATLAB Reference Guide.* Natick, MA, 1992.

[114] J. M. McNamee. Bibliography on roots of polynomials. *J. Comp. Appl. Math.*, 47, pp. 391–394, 1993.

[115] J. M. McNamee. A 2000 updated supplementary bibliography on roots of polynomials. *J. Comp. Appl. Math.*, pp. 142, 433–434, 2000.

[116] J. M. McNamee. A 2002 updated supplementary bibliography on roots of polynomials. *J. Comp. Appl. Math.*, 110, pp. 305–306, 2002.

[117] J. M. McNamee. *Numerical Methods for Roots of Polynomials. Part I.* Studies in Computational Mathematics, 14, Elsevier, 2007.

[118] J. M. McNamee, V. Pan. *Numerical Methods for Roots of Polynomials. Part II.* Studies in Computational Mathematics, 16, Elsevier, 2013.

[119] A. Melman. Modified Gershgorin disks for companion matrices. *SIAM Review*, 54, pp. 355–373, 2012.

[120] V. Mehrmann, Hongguo Xu. An analysis of the pole placement problem. I. The single-input case. *Electron. Trans. Num. Anal.*, 4, pp. 89–105, 1996.

[121] D. Mezher, B. Philippe. PAT - a reliable path-following algorithm. *Numer. Algorithms*, 29, pp. 131–152, 2002.

[122] G. S. Miminis, C.C. Paige. A direct algorithm for pole assignment of time-invariant multi-input linear systems using state feedback. *Automatica J. IFAC.*, 24, pp. 343–356, 1988.

[123] C. Moler. Cleve's corner: roots-of polynomials, that is. *The Mathworks Newsletter*, 5, pp. 8–9, 1991.

[124] R. G. Mosier. Root neighborhoods of a polynomial. *Math. Comp.*, 47, pp. 265–273, 1986.

[125] B. Mourrain, V. Y. Pan. Multivariate polynomials, duality and structured matrices. *J. Complexity*, 16(1), pp. 110–180, 2000.

[126] Y. Nakatsukasa, V. Noferini. On the stability of computing polynomial roots via confederate linearizations. *MINS EPrint*, 2014.49. UK, Manchester Institute for Mathematical Sciences, The University of Manchester.

[127] X. -M. Niu, T. Sakurai. A method for finding the zeros of polynomials using a companion matrix. *Japan J. Indust. Appl. Math.*, 20, pp. 239–256, 2003.

[128] E. E. Osborne. On pre-conditioning of matrices. *J. ACM*, 7, pp. 338–345, 1960.

[129] V. Pan. Solving a polynomial equation: history and recent progress. *SIAM Review* 39(2), pp. 187–220, 1997.

[130] B. N. Parlett, C. Reinsch. Balancing a matrix for calculation of eigenvalues and eigenvectors. *Numer. Math.*, 13, pp. 293–304, 1963.

[131] C. E. M. Pearce. On the solution of a class of algebraic matrix Riccati equation. *IEEE Trans. Automat. Control*, 31, pp. 252–255, 1986.

[132] G. Peters, J. H. Wilkinson. Practical problems arising in the solution of polynomial equations. *J. Inst. Maths. Appl.*, 8, pp. 16–35, 1971.

[133] P. Hr. Petkov, N. D. Christov, M. M. Konstantinov. A computational algorithm for pole assignment of linear multiinput systems. *IEEE Trans. Automat. Control*, 31, pp. 1044–1047, 1986.

[134] S. C. Reddy, P. J. Schmid, D. S. Henningson. Pseudospectra of the Orr-Sommerfeld operator. *SIAM J. Appl. Math.*, 53, pp. 15–47, 1993.

[135] S. C. Reddy, L. N. Trefethen. Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues. *Comput. Methods Appl. Mech. Engrg.*, 80, pp. 147–164, 1990.

[136] R. Renaut. Stability of a Chebyshev pseudospectral solution of the wave equation with absorbing boundaries. *J. Comput. Appl. Math.*, 87, pp. 243–259, 1997.

[137] K. S. Riedel. Generalized epsilon-pseudospectra. *SIAM J. Numer. Anal.*, 31, pp. 1219-1225, 1994.

[138] H. Rosenbrock. *Computer-Aided Control System Design.* Academic Press, 1974.

[139] F. Rouillier. Solving zero-dimensional systems through the rational univariate representation. *AAECC*, 9(5), pp. 443–461, 1999.

[140] P. J. Schmid, D. S. Henningson *Stability and Transition in Shear Flows.* Springer-Verlag, New York, 2001.

[141] Bl. Sendov, A. Andreev, N. Kjurkchiev. *Numerical solution of polynomial equations.* P. G. Ciarlet, J. L. Lions (Eds.), Handbook of Numerical Analysis, volume III: Solution of Equations in $\mathbb{R}^n$ (Part 2), Elsevier, Amsterdam, 1994.

[142] B. T. Smith. A zerofinding algorithm for polynomials using Laguerre's method. Technical report, Depart. of Computer Science, University of Toronto, 1967.

[143] H. P. Starr. *On the numerical solution of one-dimensional integral and differential equations.* Ph.D thesis, Yale University, Research Report YALEU/DCS/RR-888, 1992.

[144] W. W. Stewart, Ji-guang Sun. *Matrix Perturbation Theory.* Academic Press, San Diego, 1990.

[145] T. Ström. Minimization of norms and logarithmic norms by diagonal similarities. *Computing*, 10, pp. 1–7, 1972.

[146] J. G. Sun Perturbation analysis of the pole assignment problem. *SIAM J. Matrix Anal. Appl.*, 17, pp. 313–331, 1996.

[147] F. Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.*, 309, pp. 339–361, 2000.

[148] F. Tisseur, N.J. Higham. Structured pseudospectra for polynomial eigenvalue problems, with applications. *SIAM J. Matrix Anal. Appl.*, 23(1), pp. 187-208, 2001

[149] F. Tisseur, K. Meerbergen. The quadratic eigenvalue problem. *SIAM Review*, 43, pp. 235–286, 2001.

[150] K. -C. Toh, L. N. Trefethen. Pseudozeros of polynomials and pseudospectra of companion matrices. *Numer. Math.*, 68, pp. 403–425, 1994.

[151] A. Townsend, V. Noferini, Y. Nakatsukasa. Vector spaces of linearizations for matrix polynomials: a bivariate polynomial approach. *MINS EPrint*, 2012.118. UK, Manchester Institute for Mathematical Sciences, The University of Manchester.

[152] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, T. A. Driscoll. Hydrodynamic stability without eigenvalues. *Science*, 261, pp. 578–585, 1993.

[153] L. N. Trefethen, A. E. Trefethen, P. J. Schmid. Spectra and pseudospectra for pipe Poiseuille flow. *Comput. Methods Appl. Mech. Eng.*, 175, pp. 413–420, 1999.

[154] L. N. Trefethen. Computation of pseudospectra. *Acta Numerica*, pp. 247–295, 1999.

[155] L. N. Trefethen, M. Embree. *Spectra and Pseudospectra, The Behaviour of Nonnormal Matrices and Operators.* Princeton University Press, Princeton, 2005.

[156] M. Van Barel, R. Vandebril, P. Van Dooren, K. Frederix. Implicit double shift QR-algorithm for companion matrices. *Numer. Math.*, 116, pp. 177–212, 2010.

[157] P. Van Dooren. The generalized eigenstructure problem in Linear System Theory. *IEEE Trans. Automat. Control.*, AC–26, pp. 111–129, 1981.

[158] P. Van Dooren. Reducing subspaces: definitions, properties, and algorithms. *Matrix Pencils, Lecture Notes in Mathematics*, 973, B. Kågström and A. Ruhe, Eds., Springer-Verlag, Berlin, pp. 58–73, 1983.

[159] P. Van Dooren. The computation of Kronecker's canonical form of a singular pencil. *Linear Algebra Appl.*, 27, pp. 103–140, 1979.

[160] P. Van Dooren, P. Dewilde. The eigenstructure of an arbitrary polynomial matrix: computational aspects. *Linear Algebra Appl.*, 50, pp. 545–579, 1983.

[161] R. Vandebril, M. Van Barel, N. Mastronardi. *Matrix Computations and Semiseparable Matrices. Volumen 1. Linear Systems.* The John Hopkins University Press, Baltimore, 2008.

[162] J. L. M. Van Dorsselaer. Pseudospectra for matrix pencils and stability of equilibria. *BIT*, 37, pp. 833-845, 1997.

[163] A. Varga. A Schur method for pole assignment, *IEEE Trans. Automat. Control*, 26, pp. 517–519, 1981.

[164] S. Vologiannidis, E. N. Antoniou. A permuted factors approach for the linearization of polynomial matrices. *Math. Control Signal. Syst.*, 22, pp. 317–342, 2011.

[165] J. L. Walsh. On Pellet's theorem concerning the roots of a polynomial. *Ann. Math.*, 26, pp. 59–64, 1924.

[166] D. S. Watkins. A case where balancing is harmful. *Electron. Trans. Numer. Anal.*, 23, pp. 1–4, 2006.

[167] J. H. Wilkinson. The evaluation of the zeros of ill-conditioned polynomials. Part I. *Numer. Math.*, 1, pp. 150–166, 1959.

[168] J. H. Wilkinson. *Rounding Errors in Algebraic Processes.* Prentice-Hall, Englewood Cliffs, 1963.

[169] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.

[170] T. G. Wright. `EigTool`. *http://www.comlab.ox.ac.uk/pseudospectra/eigtool/*, 2002.

[171] W. M. Wonham. On pole assignment in multi-input controllable linear systems. *IEEE Trans. Automat. Control*, 12, pp. 660–665, 1967.

[172] P. Zhlobich. Differential qd algorithm with shifts for rank-structured matrices. *SIAM J. Matrix Anal. Appl.*, 33, pp. 1153–1171, 2012.